CAHIERS DU CENTAL 7

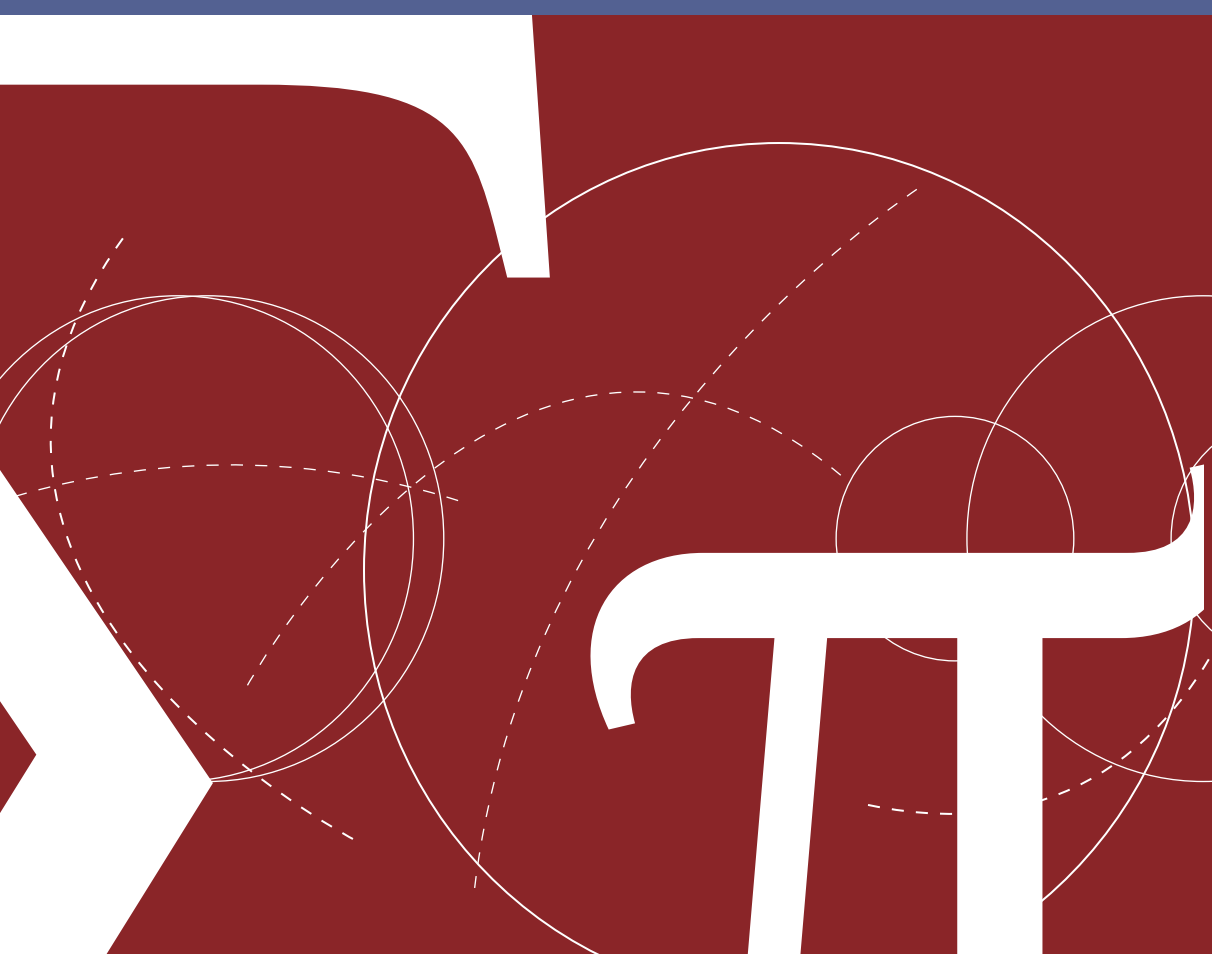
2010

eLexicography in the 21st Century: New Challenges, New Applications

Proceedings of eLex 2009, Louvain-la-Neuve, 22-24 October 2009

Sylviane Granger and Magali Paquot (eds)

UCL PRESSES UNIVERSITAIRES DE LOUVAIN



The field of lexicography is undergoing a major revolution. The rapid replacement of the traditional paper dictionary by electronic dictionaries opens up exciting possibilities but also constitutes a major challenge to the field. The eLexicography in the 21st Century: New Challenges, New Applications conference organized by the Centre for English Corpus Linguistics of the Université catholique de Louvain in October 2009 aimed to bring together the many researchers around the world who are working in the fast developing field of electronic lexicography and to act as a showcase for the latest lexicographic developments and software solutions in the field. The conference attracted both academics and industrial partners from 30 different countries who presented electronic dictionary projects dealing with no less than 22 languages.

The resulting proceedings volume bears witness to the tremendous vitality and diversity of research in the field. The volume covers a wide range of topics, including:

- *the use of language resources for lexicographic purposes, in the form of lexical databases like WordNet or corpora of different types*
- *innovative changes to the dictionary structure afforded by the electronic medium, in particular multiple access routes and efficient integration of phraseology*
- *specialised dictionaries (e.g. SMS dictionaries, sign language dictionaries)*
- *automated customisation of dictionaries in function of users' needs*
- *exploitation of Natural Language Processing tools*
- *integration of electronic dictionaries into language learning and teaching tools*

i6doc.com
L'édition universitaire en ligne

CENTAL
centre de traitement automatique du langage

ISBN 978-2-87463-211-2

eLexicography in the 21st century

New challenges, new applications

eLexicography in the 21st century

New challenges, new applications

Proceedings of eLex 2009

Louvain-la-Neuve
22-24 October 2009

Sylviane Granger and Magali Paquot (eds)

UCL PRESSES UNIVERSITAIRES DE LOUVAIN

Bernadette Dehottay CENTAL, Université catholique de Louvain, BE
Place Blaise Pascal, 1 – B-1348 Louvain-la-Neuve
tél. : +32 10 47 37 86 – fax : +32 10 47 26 06
email : cental@tedm.ucl.ac.be – http://www.uclouvain.be/cental

Membre de l'Association des Revues Scientifiques et Culturelles – A.R.S.C. (http://www.arsc.be)

Graphisme de la couverture Olivier Vereecken http://aphine.com

Dépôt légal : D/2010/9964/10
ISBN 978-2-87463-211-2
ISSN : 1783-2845

Imprimé en Belgique

Diffusion :
www.i6doc.com l'édition universitaire en ligne

Sur commande en librairie ou à :
Diffusion universitaire CIACO
Grand-Place, 7
1348 Louvain-la-Neuve, Belgique
Tél. +32 10 47 33 78 - Fax +32 10 45 73 50
duc@ciaco.com

Diffusion pour la France :
Librairie Wallonie-Bruxelles
46 rue Quincampoix
75004 Paris
Tél. +33 1 42 71 58 03 - Fax +33 1 42 71 58 09
libwabr@club-internet.fr

Table of Contents

Table of Contents v

Acknowledgements ix

Papers 1

Andrea ABEL, *Towards a systematic classification framework for dictionaries and CALL* 3

James BREEN, *Identification of neologisms in Japanese by corpus analysis* 13

Annelen BRUNNER, Kathrin STEYER, *Wortverbindungsfelder – Fields of multi-word expressions* 23

Louise-Amélie COUGNON, Richard BEAUFORT, *SSLD: a French SMS to Standard Language Dictionary* 33

Patrick DROUIN, *Extracting a bilingual transdisciplinary scientific lexicon* 43

Isabel DURÁN-MUÑOZ, *Specialised lexicographical resources: a survey of translators' needs* 55

Carolina FLINZ, *DIL: a German-Italian online specialized dictionary of linguistics* 67

Nuria GALA, Véronique REY, *Acquiring semantics from structured corpora to enrich an existing lexicon* 77

Sylviane GRANGER, Magali PAQUOT, *Customising a general EAP dictionary to meet learner needs* 87

Antton GURRUTXAGA, Igor LETURIA, Eli POCIELLO, Xabier SARALEGI, Iñaki SAN VICENTE, *Evaluation of an automatic process for specialized web corpora collection and term extraction for Basque* 97

Patrick HANKS, *Elliptical arguments: a problem in relating meaning to use* 109

Thomas HERBST, Peter UHRIG, *Valency information online – research and pedagogic reference tools* 125

Kristina HMELJAK SANGAWA, Tomaž ERJAVEC, Yoshiko KAWAMURA, *Automated collection of Japanese word usage examples from a parallel and a monolingual corpus* 137

Olga KARPOVA, Mikhail GORBUNOV, *Cultural values in a learner's dictionary: in search of a model* 149

Annette KLOSA, *On the combination of automated information and lexicographically interpreted information in two German online dictionaries* ... 157

Jette Hedegaard KRISTOFFERSEN, Thomas TROELSGÅRD, *Making a dictionary without words: lemmatization problems in a sign language dictionary* 165

Vincent LANNOY, *Free online dictionaries: why and how?* 173

Godelieve LAUREYS, *The Hub&Spoke Model put into practice. Report on the semi-automatic extraction of a pre-version of a Finnish-Danish dictionary from a multilingual interlinkable database* 183

Robert LEW, Patryk TOKAREK, *Entry menus in bilingual electronic dictionaries*... 193

Marie-Claude L'HOMME, *Designing specialized dictionaries with natural language processing. Examples of applications in the fields of computing and climate change* 203

Hanhong LI, *Word frequency distribution for electronic learner's dictionaries* ... 217

Marc LUDER, *Building an OLIF-based lexical database for representing constructions* 229

Cédric MESSIANT, Thierry POIBEAU, *Automatic lexical acquisition from corpora: some limitations and tentative solutions* 241

Mojca PECMAN, Claudie JUILLIARD, Natalie KÜBLER, Alexandra VOLANSCHI, *Processing collocations in a terminological database based on a cross-disciplinary study of scientific texts* 249

Bálint SASS, Júlia PAJZS, *FDVC – Creating a corpus-driven frequency dictionary of verb phrase constructions for Hungarian* 263

Stefania SPINA, *The Dici project: towards a dictionary of Italian collocations integrated with an online language learning platform* 273

Daniela TISCORNIA, *An ontology-based approach to the multilingual complexity of law* 283

Sabine TITTEL, *Dynamic access to a static dictionary: a lexicographical "cathedral" lives to see the twenty-first century – the Dictionnaire étymologique de l'ancien français* 295

Lars TRAP-JENSEN, *Access to multiple lexical resources at a stroke: integrating dictionary, corpus and Wordnet data* 303

Agnès TUTIN, *Showing phraseology in context: onomasiological access to lexico-grammatical patterns in corpora of French scientific writings* 313

Isabel VERDAGUER, Elisabet COMELLES, Natàlia J. LASO, Eva GIMÉNEZ, Danica SALAZAR, *SciE-Lex : an electronic lexical database for the Spanish medical community* 325

Serge VERLINDE, *The Base lexicale du français: a multi-purpose lexicographic tool* 335

Alexandra VOLANSCHI, Natalie KÜBLER, *Building an electronic combinatory dictionary as a writing aid tool for researchers in biology* 343

Michael ZOCK, Tonio WANDMACHER, Ekaterina OVCHINNIKOVA, *Are vector-based approaches a feasible solution to the "tip-of-the-tongue" problem?* 355

Posters and software demonstrations 367

Margarita ALONSO RAMOS, Alfonso NISHIKAWA, Orsolya VINCZE, *DiCE in the web. An online Spanish collocation dictionary* 369

Margarita ALONSO RAMOS, Leo WANNER, Nancy VAZQUEZ VEIGA, Orsolya VINCZE, Estela MOSQUEIRA SUAREZ, Sabela PRIETO GONZALEZ, *Tagging collocations for learners* 375

James BREEN, *WWWJDIC – A feature-rich WWW-based Japanese dictionary* 381

Elisa CORINO, Cristina ONESTI, *Have I got the wrong definition of …? How to write simple technical definitions on the basis of examples taken from Newsgroup discussions* 387

Nathalie GASIGLIA, *Some editorial orientations for a multi-tier electronic monolingual school dictionary* 393

Fadila HADOUCHE, Marie-Claude L'HOMME, Guy LAPALME, *Automatic annotation of actants in specialized corpora* 399

Jakob HALSKOV, Pia JARVAD, *Automated extraction of neologisms for lexicography* 405

Adam KILGARRIFF, Vojtěch KOVAR, Pavel RYCHLÝ, *Tickbox Lexicography* 411

Vera KUZMINA, Anna RYLOVA, *ABBYY Lingvo electronic dictionary platform and Lingvo Content dictionary writing system* 419

Margit LANGEMETS, Andres LOOPMANN, Ülle VIKS, *Dictionary management system for bilingual dictionaries* 425

Héctor MARTÍNEZ, Marta VILLEGAS, Núria BEL, Santiago BEL, Francesca ALEMANY, *Lexicography in the grid environment* 431

Carolin MÜLLER-SPITZER, Christine MÖHRS, *The "Online Bibliography of Electronic Lexicography" (OBELEX)* 439

Cornelia TSCHICHOLD, *From lexical database to intelligent vocabulary trainers* 445

Eveline WANDL-VOGT, *Multiple access routes. The dictionary of Bavarian dialects in Austria / Wörterbuch der bairischen Mundarten in Österreich (WBÖ)* 451

Christos TSALIDIS, Mavina PANTAZARA, Panagiotis MINOS, Elena MANTZARI, *NLP tools for lexicographic applications in Modern Greek)* 457

Acknowledgements

We would like to thank our academic partners and sponsors for their support of the conference.

Academic partners

- The European Association for Lexicography (EURALEX)
- The ACL Special Interest Group on the Lexicon (SIGLEX)
- Fonds National de la Recherche Scientifique
- Faculté de Philosophie, Arts et Lettres, UCLouvain
- Institut Langage et Communication, UCLouvain
- Département de Langues et Littératures Germaniques, UCLouvain

Main sponsors

- Erlandsen Media Publishing (EMP)
- Ingénierie Diffusion Multimédia (IDM)
- John Benjamins Publishing Company
- Macmillan Dictionaries Ltd
- TshwaneDJe Dictionary Production Solutions

Supporting sponsors

- ABBYY
- K Dictionaries Ltd
- Oxford University Press

Papers

Towards a systematic classification framework for dictionaries and CALL

Andrea Abel[1]
European Academy Bozen/Bolzano (EURAC)

Abstract

This paper discusses Computer-Assisted Language Learning (CALL) and dictionaries, a very specialised research field in constant evolution. A wide range of products, resources and projects are currently available or being developed, with extremely varied aims and target groups. However, the lack of in-depth analyses and of a solid classification framework calls for an attempt to systematize the subject field of dictionaries in CALL environments, which is proposed here. The outlined classification should serve as a basis for the elaboration of guidelines and evaluation parameters for such dictionaries.

Keywords: CALL, dictionaries, classification framework, design, evaluation.

1. Introduction

This paper tackles a very special niche in the field of lexicography, namely Computer-Assisted Language Learning (CALL) and dictionaries. While there is a longstanding tradition in lexicographic as well as metalexicographic research, and a lot about CALL has been written during the last decades, there is little written specifically about dictionaries in CALL systems.

Currently, a wide range of different products and projects are available or being developed, and there is a wide range of different resources, aims and target groups. In this regard, a lack of a systematic and sound analysis as well as a solid classification framework can be noticed. This paper therefore aims at filling this gap and at providing an attempt at a first systematisation of dictionaries in CALL environments as a basis for the elaboration of guidelines and evaluation strategies for them.

2. CALL and dictionaries – some definitions

The use of new technologies and media for language learning and teaching has become a separate discipline known as CALL. CALL stands for Computer-Assisted

[1] Institute for Specialised Communication and Multilingualism, EURAC, andrea.abel@eurac.edu

Language Learning (or Computer-Aided Language Learning) and is defined by Levy (1997: 1) as

> the search for and study of applications of the computer in language teaching and learning.

Within the field of CALL a number of terms and acronyms are used, *e.g.* CELL (Computer-Enhanced Language Learning), TELL (Technology-Enhanced Language Learning), ICALL (Intelligent Computer-Assisted Instruction), NCALL (Network based CALL) (*cf.* Levy 1997: 80; Knapp 2004). Each term places the focus on a particular aspect of the field or reflects a certain trend.

The term "dictionary" does not need to be defined here, as the target group of this publication is assumed to be subject field experts. Nevertheless, in the following paragraphs some terms will be briefly described in relation to the meaning they are used in within the present paper.

In this context the term "dictionary" refers to electronic dictionaries. Defining what an "electronic dictionary" is, is indeed not a trivial task. A generic description, which can be used to convey the main idea, is given by Nesi (2000: 839):

> The term electronic dictionary (or ED) can be used to refer to any reference material stored in electronic form that gives information about spelling, meaning, or use of words. Thus, a spell-checker in a word-processing program, a device that scans and translates printed words, a glossary for on-line teaching materials, or an electronic version of a respected hard-copy dictionary are all EDs of a sort […].

Furthermore, de Schryver (2003: 146) describes EDs as

> collections of structured electronic data that can be accessed with multiple tools, enhanced with a wide range of functionalities, and uses in various environments.

Turning now to CALL and dictionaries, there is an obvious link to the concept of learners' dictionary, which can be generally defined as

> a dictionary whose genuine purpose is to satisfy the lexicographically relevant information needs that learners may have in a range of situations in connection with the foreign-language learning process (Tarp 2008: 130).

However, requirements for dictionaries in CALL-environments can differ from those for learners' dictionaries, because they are strongly connected to the overall aim and approach of the whole CALL application.

Before moving on to the description of the proposed classification framework, some more specifications are needed: in this paper the term "CALL" refers only to applications including a dictionary component. As to "lexicography", in this special context it is used in relation to the use of dictionaries in CALL.

3. Towards a classification framework

At the present stage the proposed classification for dictionaries and CALL basically moves along some central dimensions, but can easily be extended further and, already

in this form, serves as a first basis for analysing and evaluating existing applications in a more systematic way than in the past.

These central dimensions are described in the following paragraphs and briefly outlined with some concrete examples.

3.1. CALL-cum-dictionary *vs* dictionary-cum-CALL-systems

The first dimension can best be captured by the notions of CALL-cum-dictionary-systems *vs* dictionary-cum-CALL-systems, by analogy with the concept of dictionary-cum-corpus-systems, going back to a formulation by Leech (1997). The term "CALL-cum-dictionary" emphasizes the central role of the CALL system where the dictionary has been added as an aid, as one among others. Such systems probably account for the majority of CALL systems.

In contrast, the notion of dictionary-cum-CALL is of particular importance in a primarily lexicographic context, as the dictionary is the central element and/or the starting point for the whole CALL application.

For the CALL-cum-dictionary-systems many examples could be mentioned. A small but highly diversified selection is listed here; the focus is not on commercial software, but on products developed in public or in research environments: *e.g.* ESPRIT is conceived as an interactive plurilingual ICALL software system for autonomous, contrastive learning of French, Spanish and Italian (Koller 2005; Koller 2007); whereas WUFUN is a multimedia computer-assisted tutoring system for vocabulary learning which has been developed for Chinese university learners of English (Ma and Kelly 2006); LINC is a multimedia CALL-application on CD-Rom for different European languages (Poel and Swanepoel 2003); the Compleat Lexical Tutor website contains a multiplicity of online resources for both teaching and learning vocabulary and grammar (Cobb 2007; Horst, Cobb and Nicolae 2005; Sevier 2004); the computer research prototype tool DEFI is intended to rank translations according to their relevance in a given context (Michiels 2000); Glosser-RuG (Nerbonne and Smit 1996)/Glosser-WeB (Dokter 1997) is a CALL tool which is aimed at assisting French readers in developing their comprehension and reading skills; Deutsch-Uni Online is a commercial language learning platform aimed at people who want to study or work in Germany (Roche 2007); "Tell me more" has been chosen among others as a typical sample of a commercial language learning software on CD-Rom (Auralog, 2005, Spanish Level 3).

Finally, some examples serve to illustrate the notion of dictionary-cum-CALL-systems. On the one hand, dictionaries can be the basis upon which exercises related to dictionary components can be built. These components can be used either in simple pattern-drill-based vocabulary trainers (*e.g.* Pons Lexitrainer, http://www.lexitrainer.de/) or in intelligent environments, *e.g.* BLF-ALFALEX-DAFLES, an online lexical database for learners of French with special focus on vocabulary learning (Verlinde, Selva and Binon 2006), or the ELDIT-program, a crosslingual

learners' dictionary for German and Italian (Abel and Weber 2000; Knapp 2004) as well as a system aiming at web-based language learning (Knapp 2004; Knapp and Höber 2006). On the other hand, dictionary creation itself can be the aim of the system through which language learning, especially vocabulary and collocation learning, can be supported, *e.g.* LogoTax (Ludewig 2005).

3.2. Human-oriented *vs* machine-oriented dictionaries/lexicons and CALL

A second dimension of dictionaries used in CALL is the distinction between human-oriented applications ("dictionaries" in a more restricted acceptation of the term) *vs* machine-oriented tools (in the following referred to as "lexicons") (*cf. e.g.* de Schryver 2003: 144-146 on the discussion of the terminology). Human-oriented dictionaries are those that are usable through a graphical user interface (GUI), and the interface design is crucial. However, machine-oriented lexicons are based on a complex internal representation that is accessed by the system, *e.g.* for facilitating dictionary lookup, error diagnosis and error feedback. Dictionaries can also contain both elements. Hence, this can be interpreted as a new facet of the multifunctionality of a dictionary (*cf.* Heid and Gouws 2006: 981 on the notion of multifunctionality, Abel 2003: 537 on the flexibility and modularity of new electronic dictionaries as a basis for including them in multiple environments).

Some examples:

With reference to human-oriented dictionaries, learners' dictionaries such as the ones produced by Longman and Cambridge, for instance, are used within several tools of the Compleat Lexical Tutor Websites (*e.g.* Horst, Cobb and Nicolae 2005).

Machine-oriented lexicons are used in different applications, too, so for instance the "e-assistant" in Deutsch-Uni Online, which consists of a spelling and grammar checker and gives error feedback to the learner (Roche 2007); another example is the MIRTO toolbox for language learning, which uses NLP resources for morphosyntactic tagging etc. (Antoniadis *et al.* 2004).

Finally, some applications include both, human-oriented dictionaries as well as machine-oriented lexicons, *e.g.* the ELDIT dictionary (Knapp 2004) or the reading aid Glosser (Nerbonne and Smit 1996; Dokter 1997).

3.3. Primarily *vs* secondarily CALL-oriented or lexicography-oriented applications

A third dimension is the distinction between applications that are primarily *vs* secondarily CALL-oriented or lexicography-oriented. This aspect could be especially interesting for future research.

While CALL applications are obviously systems developed especially for language learning, there are also systems, which are inherently intended for other purposes, but can at the same time be used for language learning. The latter shall be referred to as "secondarily CALL-oriented applications" in contrast to "primarily CALL-oriented applications". Some applications are especially developed for lexicographic purposes,

while others are not, *i.e.* "primarily lexicography-oriented applications" as distinguished from "secondarily lexicography-oriented applications".

Starting from this basis four different combinations can be identified, which are briefly exemplified in the following paragraphs.

The most obvious or "normal" combination is that of applications that are primarily CALL-oriented and secondarily lexicography-oriented: *e.g.* commercial products, such as "Tell me more", programs like LINC (Poel and Swanepoel 2003) or LingQ, an online language learning platform for a variety of languages with a focus on vocabulary learning (http://www.lingq.com/), but also reading aids, such as Glosser (Nerbonne and Smit 1996; Dokter 1997). Within these applications different language skills are trained and dictionaries are added as an aid.

In some systems language learning and lexicography are equally important; according to the above mentioned distinctions these can be called primarily CALL-oriented and primarily lexicography-oriented applications. One example is LogoTax, the tool for vocabulary learning which foresees dictionary creation (Ludewig 2005).

In addition, some systems that are inherently intended for other purposes can be used for language learning. Therefore, they can be labelled secondarily CALL-oriented and primarily lexicography-oriented applications. The research tools that can be named as examples of this category are partly experimental, partly fully operational systems, *e.g.* "Words in your ear" for the investigation of dictionary lookup patterns (Laufer and Hill 2000), or the analysis of the effects of multimedia annotation on vocabulary acquisition as described by Chun and Plass 1996. Other programs, such as collaborative dictionary writing systems, can be used for language learning, too (*e.g.* the wiki-based Open Content dictionary Wiktionary, http://www.wiktionary.org/).

Finally, some applications may be interesting for in-depth research even though they were not intended either for lexicography or for CALL, *i.e.* secondarily CALL-oriented and secondarily lexicography-oriented applications; nevertheless, they could be fruitfully used for both, as well as for other research purposes, too. As an example for this category the ESP games could be mentioned (*e.g.* von Ahn 2006), where players have to label pictures and, by doing so, they train their ability to define terms online. This can be effectively adopted for language learning purposes. Furthermore, results can be employed for lexicographic purposes, too, *e.g.* for experiments aimed at improving the formulation of definitions.

3.4. Degree of professionalism in dictionary production

Another dimension to be considered concerns the way dictionaries are produced and their degree of professionalism.

A kind of amateur lexicography (for the term *cf.* Ludewig 2005: 210) exists, which can take different forms. Collaborative lexicography (for the term *cf. e.g.* Storrer 1998: 125), where everyone can add a word entry by using a simple online form, is quite common, such as the Open Content dictionary Wiktionary (http://www.

wiktionary.org/), or smaller applications like the Word Bank tool of the Compleat Lexical Tutor (Horst, Cobb and Nicolae 2005). In contrast to this, a controlled wiki approach can be applied, with a workflow for learner-tutor-interaction (a prototype of this kind of approach is described *e.g.* in Abel and Bracco 2006). An example of amateur lexicography using professional tools without any tutor interaction is the LogoTax system (Ludewig 2005). Some projects mainly re-use already existing lexicographic data in a basically technology-driven way, without explicitly applying any specific lexicographic approach, *e.g.* popular and widely used dictionaries such as LEO (www.leo.org) or Beolingus (http://dict.tu-chemnitz.de/). Moreover, many dictionaries, glossaries, word lists etc., as well as partly abridged versions of commercial products, mostly free of charge, are widespread where no clear information on their sources is available (*e.g.* some freely available resources at www.travlang.com, www.babylon.com).

Beyond this level there is the huge field of professional lexicography producing electronic dictionaries, in many cases on CD-ROMs. In addition, more and more publishing houses offer their products or parts of them online and free of charge, next to their print products.

Finally, there is a niche of lexicographically-oriented research projects, some of which are prototypes or contain a small amount of data, because financing in many cases is a challenge (*e.g.* BLF-ALFALEX-DAFLES – Verlinde, Selva and Binon 2006; ELDIT – Abel and Weber 2000; Knapp 2004).

4. Conclusions and outlook

In conclusion, it can be stated that the requirements for dictionaries and CALL depend on the overall aim and approach of the whole application. Therefore, in future it will be interesting to further examine already existing systems, hoping to draft more specific guidelines on how to describe the crucial interdependence between system, scope and dictionary.

As a next step it would be important to further elaborate the framework and prepare a complex and flexible checklist, including a wider range of dimensions (*e.g.* regarding the existence of an explicit design statement, the use of specific resources, the inclusion of NLP-techniques, the human-computer interaction, the kind and level of integration of electronic dictionaries in CALL), for the evaluation of existing as well as for the design and the creation of new applications.

References

ABEL, A. (2003). *Alte und neue Problematiken der Lernerlexikographie in Theorie und Praxis*. Ph.D. (unpublished). Innsbruck.

ABEL, A. and BRACCO, St. (2006). From an online dictionary to an online dictionary writing system. In G.-M. de Schryver (ed.). *DWS 2006: Proceedings of the Fourth International*

Workshop on Dictionary Writing Systems. Pretoria: $(SF)^2$ Press: 25-34 (http://nlp.fi.muni.cz/dws06/dws2006.pdf).

ABEL, A. and WEBER, V. (2000). ELDIT – A Prototype of an Innovative Dictionary. In U. Heid, S. Evert, E. Lehmann and C. Rohrer (eds). *Proceedings of the Ninth International Congress, EURALEX 2000, Stuttgart, Germany, August 8th-12th.* Stuttgart: Institut für Maschinelle Sprachverarbeitung, vol. II: 807-818.

ANTONIADIS, G., ECHINARD S., KRAIF, O. *et al.* (2004). NLP-based scripting for CALL activities. In L. Lemnitzer, D. Meurers and E. Hinrichs (eds). *Proceedings of the Workshop on eLearning for Computational Linguistics and Computational Linguistics for eLearning (COLING - The 20th International Conference on Computational Linguistics).* Geneva: Association for Computational Linguistics: 18-25.

Babylon: www.babylon.com

Beolingus: http://dict.tu-chemnitz.de/

CHUN, D.M. and PLASS, J.L. (1996). Effects on multimedia annotations on vocabulary acquisition. *The Modern Language Journal*, 80(2): 183-198.

COBB, T. (2007). Computing the Vocabulary Demands of L2 Reading. *Language Learning and Technology*, 9(2): 90-110.

DE SCHRYVER, M.-G. (2003). Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography*, 16(2): 143-199.

DOKTER, D. (1997). From Glosser-RuG to Glosser-WeB. *Technical Report* (http://www.let.rug.nl/glosser/).

HEID, U. and GOUWS, R.H. (2006). A Model for a Multifunctional Dictionary of Collocations. In E. Corino, C. Marello and Cr. Onesti (eds). *Atti del XXII Congresso Internazionale di Lessicografia. Torino, 6-9 settembre 2006 (=Proceedings XII EURALEX International Congress).* Alessandria: Edizioni dell'Orso, vol. 1: 979-988.

HORST, M., COBB, T. and NICOLAE, I. (2005). Expanding Academic Vocabulary with an Interactive On-Line Database. *Language Learning and Technology*, 9(2): 90-110.

KNAPP, J. (2004). *A New Approach to CALL Content Authoring*. Ph.D. Hannover (http://www.eurac.edu/NR/rdonlyres/83E0D545-899B-45A7-B144-281140FB9B9E/0/knappPhD.pdf).

KNAPP, J. and HÖBER, A. (2006). Quantity OR quality – an inevitable dilemma in innovative CALL systems? In A. Abel, M. Stuflesser and M. Putz (eds). *Mehrsprachigkeit in Europa: Erfahrungen, Bedürfnisse, Gute Praxis. Tagungsband. - Plurilinguismo in Europa: esperienze, esigenze, buone pratiche. Atti del convegno. - Multilingualism across Europe: Findings, Needs, Best Practices. Proceedings. 24.-26.08.2006*, Bolzano/Bozen. Bozen: Eurac: 126-140. (http://www.eurac.edu/NR/rdonlyres/E4DF9F14-FE8C-4E44-A221-C0836D96C428/0/Multilingualismindb.pdf).

KOLLER, Th. (2005). Development of web-based plurilingual learning software for French, Spanish and Italian. In C. Mourón Figueroa and T.I. Moralejo Gárate (eds). *Studies in Contrastive Linguistics. Proceedings of the 4th International Contrastive Linguistics Conference (ICLC4)*. Santiago de Compostela, University of Santiago de Compostela Press: 461-469.

KOLLER, Th. (2007). *Design, Development, Implementation and Evaluation of a Plurilingual ICALL System for Romance Languages Aimed at Advanced Learners*. Ph.D. Dublin.

LAUFER, B. and HILL, M. (2000). What lexical information do L2 learners select in a CALL dictionary and how does it affect word retention. *Computer Assisted Language Learning (CALL)*, 3(2). Special Issue: The Role of Computer Technology in Second Language Acquisition Research: 58-76.

LEECH, G. (1997). Teaching and Language Corpora: a Convergence. In A. Wichmann, St. Fligelstone and T. McEnery (eds). *Teaching and Language Corpora*. London: Addison-Wesley Longman: 1-23.

Leo: www.leo.org

LEVY, M. (1997). *Computer-Assisted Language Learning: Context and Conceptualization*. Oxford: Clarendon Press.

LingQ: http://www.lingq.com/

LUDEWIG, P. (2005). *Korpusbasiertes Kollokationslernen*. Frankfurt a.M: Peter Lang.

MA, Q. and KELLY, P. (2006). Computer Assisted Vocabulary Learning: Design and Evaluation. *Computer Assisted Language Learning (CALL)*, 19(1): 15-45.

MICHIELS, A. (2000). New developments in the DEFI Matcher. *International Journal of Lexicography*, 13(3): 151-167.

NERBONNE, J. and SMIT, P. (1996): GLOSSER-RuG: in Support of Reading. In *COLING 06. The 16th International Conference on Computational Linguistics*. Groningen. vol. 2: 830-836.

NESI, H. (2000). Electronic Dictionaries in Second Language Vocabulary Comprehension and Acquisition: the State of the Art. In U. Heid, S. Evert, E. Lehmann and C. Rohrer (eds). *Proceedings of the Ninth International Congress, EURALEX 2000, Stuttgart, Germany, August 8th-12th.* Stuttgart: Institut für Maschinelle Sprachverarbeitung: 839-847.

POEL, K. and SWANEPOEL, P. (2003). Theoretical and Methodological Pluralism in Designing Effective Lexical Support for CALL. *Computer Assisted Language Learning (CALL)*, 16(2): 173-211.

PONS LEXITRAINER: http://www.lexitrainer.de/

ROCHE, J. (ed.) (2007). *Benutzerhandbuch für die Arbeit mit den Online-Kursmodulen der Deutsch-Uni Online*. München. (http://www.uni-deutsch.de/help/pdfs/DUO-Benutzerhandbuch.pdf)

SEVIER, M. (2004). The Compleat Lexical Tutor (v.4). (Review). *Teaching English as a Second or Foreign Language*, 8(3), (http://www-writing.berkeley.edu/TESL-EJ/ej31/m2.html).

STORRER, A. (1998). Hypermedia-Wörterbücher: Perspektiven für eine neue Generation elektronischer Wörterbücher. In E.H. Wiegand (ed.). *Wörterbücher in der Diskussion III*. Tübingen: Niemeyer (*Lexicographica: Series Maior*).

TARP, S. (2008). *Lexicography in the Borderland between Knowledge and Non-Knowledge. General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Tübingen: Niemeyer (*Lexicographica: Series Maior*).

Tell me more, Spanish Level 3. Auralog, 2005.

Travlang: www.travlang.com

VERLINDE, S., SELVA, Th. and BINON, J. (2006). The Base lexicale du français (BLF): A Multifunctional Online Database for Learners of English. In E. Corino, C. Marello and Cr. Onesti (eds). *Atti del XXII Congresso Internazionale di Lessicografia. Torino, 6-9 settembre 2006 (= Proceedings XII EURALEX International Congress).* Alessandria: Edizioni dell'Orso. vol. 1: 471-481.

VON AHN, L. (2006). Games with a purpose. *IEEE Computer Magazine*, June 2006: 96-98. (http://www.cs.cmu.edu/~biglou/ieee-gwap.pdf).

Wiktionary: http://www.wiktionary.org/

(All links checked on 07.12.2009)

Identification of neologisms in Japanese by corpus analysis

James Breen[1]
Monash University, Australia

Abstract

In Japanese and other languages that do not use spaces or other markers between words, the identification and extraction of neologisms and other unrecorded words presents some particular challenges. In this paper we discuss the problems encountered with neologism identification and describe and discuss some of the methods that have been employed to overcome these problems.

Keywords: Japanese, neologism, kanji, hiragana, katakana, segmentation, corpus, n-gram.

1. Introduction

In "The Oxford Guide to Practical Lexicography", we find the unqualified statement "It's easy for computer programs to spot completely new words" (Atkins and Rundell 2008: 51). The authors must have been thinking of European languages, where modern orthographical practice has each word separated by spaces. The quoted statement clearly does not apply to languages such as Japanese or Chinese where apart from punctuation there is no clear marker between words, and where the very concept of "word" is often debated.

In this paper we describe recent and planned work to extend some techniques reported earlier to identify and extract neologisms from Japanese texts (Breen 2004a; Breen 2005; Kaji, Uno and Katsuregawa 2009). The purpose of the research is to extend the recorded lexicon of Japanese, both in free and commercial dictionaries.

2. Overview of Japanese orthography

Modern Japanese is written in a mixture of scripts:

a) *kanji* (Chinese characters), which are used mainly for nouns and the roots of verbs, adjectives, etc. Approximately 2,000 *kanji* are in common use, although

[1] jimbreen@gmail.com

the full set available is estimated to be around 80,000. Most nouns written with *kanji* use two or more characters, whereas verbs typically use a single *kanji.*

b) the *hiragana* syllabary (46 symbols plus diacritics: あいうえおかきくけこ, etc.) In modern Japanese *hiragana* is used mainly for particles, verb and adjective inflections, conjunctions, etc.

c) the *katakana* syllabary (also 46 symbols plus diacritics: アイウエオカキクケコ, etc.) *Katakana* are currently used for loanwords, scientific names, transcriptions of foreign names, etc.

An illustration of the use of the scripts can be seen in the sentence スーパーで食品を買いました *su-pa- de shokuhin o kaimashita*: [I] bought some food at [a/the] supermarket). Here *kanji* are used for the noun 食品 (foodstuffs) and for the root of the verb 買う (*kau* to buy), *hiragana* are used for the particles で and を and the polite past-tense inflection of the verb (いました), and *katakana* is used for the abbreviated form of the loanword スーパーマーケット (*su-pa-ma-ketto* supermarket).

3. Neologisms in Japanese

Despite having a rich lexicon, the Japanese language has a noted tendency to adopt and create new words (Lee 2002; Tsujimura 2006;). While the reasons for adopting new words are varied, there are a number of processes associated with the Japanese language which tend to encourage neologism creation:

a) the readiness to accept loanwords. Unlike some countries, which attempt to restrict loanword usage, Japan has placed no formal restriction on their use. Estimates of the number of loanwords used in Japanese range as high as 80,000. Most of these words have been borrowed directly from English, however a significant number, known as *wasei eigo* (Japanese-made English) have been assembled from English words or word fragments.

b) the accepted morphological process of creating words by combining two or more *kanji* (Chinese characters) chosen for their semantic properties. This process was used extensively in the mid-19th century when Japan re-engaged with the rest of the world and needed an expanded lexicon to handle the technological, cultural, etc. information flowing into the country. This process has continued. A broadly similar process is used to create compound verbs.

c) the tendency to create abbreviations, particularly from compound nouns and long loanwords. For example, the formal term for “student discount” in Japanese is *gakusei waribiki* (学生割引), however the common term is *gakuwari* (学割) formed from the first *kanji* in each of the two constituent nouns. A similar process is applied to loanwords, resulting in words such as *sekuhara* (セクハラ) for “sexual harassment” (a contraction of *sekushuaru harasumento*).

Many neologisms find their way eventually into published dictionaries, and there are several special neologism dictionaries (*shingo jiten*, *gendaiyôgo jiten*). However, many abbreviations, compound verbs and loanwords are less well lexicalized as native speakers can usually recognize them as such and recognize the pronunciation and meaning.

Traditional techniques for identifying neologisms involve extracting lexemes and comparing them with a lexical database. This process can have problems in Japanese as the orthography does not use any separators between words. As described below, text segmentation software packages for Japanese typically use extensive lexicons to enable word segments to be identified, but have proved to be be unpredictable when out-of-lexicon strings are encountered.

4. Word segmentation in Japanese

Computerized segmentation of Japanese text was once considered a very difficult task; some writers in the 1980s thought it impossible. Since the 1990s several good systems have emerged, *e.g.* the open-source research-oriented Juman (Kyoto University) and Chasen and MeCab (Nara Institute of Science and Technology), and commercial and in-house systems from Basis Technology, NTT and Google. All of these combine artificial intelligence techniques with large lexicons (which implies that for correct operation the words must be known already).

As an example of such segmentation software consider the sentence “その教師は講堂に学生を集めた。” (that teacher assembled the students in the auditorium) when processed by the Chasen system. Table 1 shows the results of the segmentation.

The sentence has been correctly segmented, and the 集めた has been correctly identified as the た (past tense) inflection of 集める.

Such segmentation software usually outputs unassociated strings of characters when words are encountered which are not in their lexicons. In Table 2 we illustrate this by substituting some unknown words (全堂 instead of 講堂 and 凧黽 instead of 学生) which results in the following segmentation.

The 全堂 has been identified as a prefix-noun combination, which is plausible, but the *kanji* in 凧黽 have been flagged 未知語 (*michigo:* unknown word). The tendency of these software systems to output unassociated strings of characters when words are encountered which are not in their lexicons is well known. Some work has been carried out on reconstructing these “unknown words”, but usually in the context of part-of-speech tagging and dependency analysis (Asahara and Matsumoto 2004; Uchimoto, Sekine and Isahara 2001; Utsuro, Shime, Tsuchiya, Matsuyoshi and Sato 2007).

Word Segment	Reading	Lexical Form	POS Information
その	ソノ	その	連体詞
教師	キョウシ	教師	名詞-一般
は	ハ	は	助詞-係助詞
講堂	コウドウ	講堂	名詞-一般
に	ニ	に	助詞-格助詞-一般
学生	ガクセイ	学生	名詞-一般
を	ヲ	を	助詞-格助詞-一般
集め	アツメ	集める	動詞-自立 一段 連用形
た	タ	た	助動詞 特殊・タ 基本形
。	。	。	記号-句点

Table 1. Example of CHASEN Text Segmentation

Word Segment	Reading	Lexical Form	POS Information
全	ゼン	全	接頭詞-名詞接続
堂	ドウ	堂	名詞-一般
胤	-	-	未知語
毘	-	-	未知語

Table 2. Example of CHASEN Parsing Unknown Words

5. Approaches to finding new words in Japanese texts

Three broad approaches are proposed for identifying neologisms and other unlexicalized Japanese words:

a) scanning texts and other corpora for possible "new" words, typically by processing the texts through segmentation software and dealing with the "out-of-lexicon" problem;

b) mimicking Japanese morphological processes to generate possible words, then testing for the presence of the "words" in corpora;

c) application of machine learning techniques in which software has been trained to identify the language constructs typically associated with the introduction and discussion of new or rare words.

These approaches are discussed in more detail below.

6. Scanning texts for neologisms and unlexicalized words

The general approach is as follows:

a) process texts through segmentation software to extract lexemes. Ideally the lexicons used by the software should be extended to include as many known words as possible;

b) detect and analyze the cases where the analysis has failed. This will involve considerable post-processing, including careful profiling of any identified affixes, as Japanese is an agglutinative language which makes considerable use of highly productive single-character affixes;

c) extraction of possible unrecorded words;

d) examination of the words in the original textual contexts;

e) development of the reading (*i.e.* pronunciation) and the meaning of the words.

As reported in Breen (2005), an initial trial of this method was carried out in which 500 articles from the Asahi Shimbun newspaper were analyzed. The process concentrated on isolated unlexicalized *kanji* pairs. A number of hitherto unrecorded words were identified, *e.g.*

a) previously unrecorded names *e.g.* 武示 (Takeshi), 晃毅 (Kouki), 潔重 (Yukishige);

b) newly-arrived terms, *e.g.* 米紙 (American press/newspapers) and 軍歴 (military service record)

c) many abbreviations, *e.g.* 日歯連 (from 日本歯科医師連盟 - Japan Dentists Federation)

d) newspaper-style formations such as 中韓 (Chinese-Korean) and 仏誌 (French publication)

e) several apparently new formations such as 入境 (border crossing or border entry) and 公助 (public assistance)

We can draw on the fact that loanwords in Japanese are written in the *katakana* syllabary, thus enabling relatively straightforward extraction and comparison. The study also harvested unrecorded words written in *katakana.* Approximately 20% of the words in *katakana* were "new", and contained:

a) many transcribed names (esp. Chinese and Korean);

b) Japanese flora/fauna terms;

c) many variants of common loanwords *e.g.* プロフィル (profile) instead of the more common プロフィール.

d) a number of words and expressions worth adding to the lexicon, *e.g.* ピープルパワー (people-power) and ゼロメートル (zero metre, which in Japanese means sea level).

7. Generation of possible words

In this approach we mimic Japanese morphological processes to synthesize potential words, then test if the "word" exists in the lexicon or is in use in corpora.

Early trials used the WWW as a test corpus, with accesses via a programmed interface to a search engine (in this case the Google API.) A new WWW-derived resource for such testing is the Google Japanese Web N-gram Corpus (Kudo and Kazawa 2007). This corpus uses text extracted from a one-month WWW snapshot taken in July 2007. Text strings were processed through MeCab, and all 1-gram to 7-gram sequences occurring more than 20 times were counted and recorded. The resulting n-grams are published as a set of files containing from 2.5M 1-grams to 570M 7-grams (over 1.7M of the 1-grams are *katakana* words or compounds). This corpus has huge potential in corpus linguistics research and will be a very important resource in neologism detection and extraction.

A trial of the technique was carried out using synthesized *kanji* abbreviations based on the above-mentioned 4-*kanji* to 2-*kanji* pattern (*e.g.* 学生割引 being abbreviated to 学割) (Breen 2004a). Approximately 8,000 4-*kanji* compound verbs were extracted from the JMdict lexicon (Breen 2004b), 2-*kanji* abbreviations were created, and those that were not already lexicalized were tested against WWW pages. As *kanji* pairs can occur in many contexts, the text in which the potential abbreviations appeared was analyzed and classified according to the location of the *kanji* pair, surrounding kana, kanji, punctuation, etc.) and WWW page hits. Approximately 700 potential abbreviations were identified for deeper analysis, and a large number of abbreviations established.

A further study was carried out using synthesized compound verbs (Breen and Baldwin 2009). In Japanese compound verbs, formed from two (or more) verbs and acting as a single verb, are very common and highly productive. For example 歌い始める (to start singing) is formed from 歌う (to sing) and 始める (to start or begin). 2,900 compound verbs were selected from the JMdict lexicon, the two verb

portions extracted (700 and 600 respectively) and 420,000 potential compound verbs generated. These were tested in the three most common inflections against the Google n-gram corpus, and approximately 22,800 were found to be in use (of these 4,800 were recorded in a range of lexicons). Samples of the 22,800 were examined in detail, indicating that over 90% precision was being achieved.

8. Direct scan of the N-gram corpus

The availability of the Japanese n-gram corpus has opened the possibility of searching it directly for unlexicalized words. For example, with 2,000 common *kanji* the possible 2-*kanji* compounds are only 4 million, and it is possible to scan the n-gram corpus for occurrences of such compounds in suitable textual contexts such as *kana-kanji-kanji-kana* sequences.

A direct scan approach was also used as an extension of the compound verb extraction mentioned above. The n-gram corpus was scanned using a symbolic template of a compound verb and the selected candidates filtered for valid inflectional values. Approximately 80,000 possible compound verbs were detected (of which 6,200 were in the range of lexicons), and sampling indicated that a precision of approximately 60% was achieved.

9. Machine learning

As noted above, the Japanese language has a tendency to adopt and create new words. As a result, there is considerable discussion of new words in Japanese newspapers, WWW pages, etc., and there are several WWW sites in Japan devoted to such discussions. Discussion of word meanings associated with neologisms, etc. tend to follow particular linguistic patterns, for example a passage discussing the neologism オタ芸 has

“オタ芸（オタげい・ヲタげい）とは、アイドルや声優などのコンサートや ”.

The pronunciation is parenthesized after the word, and followed by the “とは” particle which is typically used to flag an explication of a term. There are a number of such linguistic patterns, and research is under way to train text classification software to detect documents containing such passages, this enabling a focussed analysis on documents likely to contain neologisms.

10. Derivation of readings

The pronunciation or reading of an unrecorded word is a function of the *kanji* with which it is written while words written in *hiragana* or *katakana* have established pronunciations. There are two issues that need to be dealt with in establishing the pronunciation of such words:

a) unlike Chinese, where each character typically has a single pronunciation, Japanese usually has several pronunciations for each character. Some pronunciations are more common than others and it is possible to generate most probable pronunciations for later testing;

b) a number of character pronunciations are not voiced when occurring at the start of a word, but voiced within a word, *e.g.* 所 is *tokoro* in initial positions, but usually *dokoro* elsewhere. The rules for this process are complex and not complete, for example 島 (island) is pronounced both *shima* and *jima* in identical contexts.

There is a tendency to write the pronunciation of unusual words in parentheses after its first occurrence in a text. This enables testing of candidate pronunciations by search for a collocation of a word with its possible pronunciation.

11. Derivation of meanings

This is, of course, traditionally the most intensive and time-consuming part of lexicography. In our corpus-based processes we have been working towards automatic derivation of candidate meanings followed by human checking and verification.

With regard to automatic derivation of meanings, some general observations can be made:

a) for abbreviations the process is relatively straightforward as the abbreviation almost always carries the meaning of the source word or expression;

b) for compound verbs there has been considerable success combining semantic and lexical information associated with the component verbs. In this area the English n-gram corpus is also proving useful in identifying most likely candidates;

c) for multi-word expressions it is often possible to get good results by testing combinations of the meanings of constituent words, *e.g.* 海底電線 → (undersea, submarine) (electric, telephone, line, cable, wire) leading to “undersea cable” or “submarine cable” as the most likely candidate;

d) loanwords written in *katakana* can be a challenge. While there is some success in back-translation into English, especially when combined with checking for collocations on WWW pages, there is a persistent problem with pseudo-loanwords constructed from foreign words or word-fragments, and from non-English loanwords (Korean, French, German, etc.);

e) compound loanword nouns/expressions can be handled similarly to others. *e.g.* スパイスライス could be parsed as spice+rice or spy+slice. However, checking against English n-grams indicates that the former is the correct translation.

12. Conclusion

The nature of Japanese orthography makes neologism detection more difficult than in many other languages.

Modern computational linguistics has techniques and resources to assist in both identification of Japanese neologisms and other unrecorded words, and in deriving readings and meanings. This is a major research area, and a lot more work remains to be done.

References

ASAHARA, M. and MATSUMOTO, Y. (2004). Japanese Unknown Word Identification by Character-based Chunking. In *International Conference on Computational Linguistics* (*COLING) 2004*, Geneva: 459-466.

ATKINS, S. and RUNDELL, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

BREEN, J. (2004a). Expanding the Lexicon: the Search for Abbreviations. In *Papillon Multi-lingual Dictionary Project Workshop*, Grenoble.

BREEN, J. (2004b). JMdict: a Japanese-Multilingual Dictionary. In *International Conference on Computational Linguistics* (*COLING) Multilingual Linguistic Resources Workshop*, Geneva: 71-78.

BREEN, J. (2005). Expanding the Lexicon: Harvesting Neologisms in Japanese. In *Papillon Multi-lingual Dictionary Project Workshop*, Chiang Rai.

BREEN, J. and BALDWIN, T. (2009). Corpus-based Extraction of Japanese Compound Verbs. In *Australasian Language Technology Workshop (ALTW2009)*, Sydney.

KAJI N., UNO, N. and KITSUREGAWA, M. (2009). Mining Neologisms from a Large Diachronic Web Archive for Supporting Linguistic Research. In *Data Engineering and Information Management (DEIM2009)*, Tokyo (in Japanese).

KUDO, T. and KAZAWA, H. (2007). *Japanese Web N-gram Corpus Version 1*, Google/Linguistic Data Consortium, http://www.ldc.upenn.edu/

LEE, S.C. (2002). Lexical Neologisms in Japanese. In *Australian Association for Research in Education Conference*, Brisbane.

TSUJIMURA, N. (2006). *An Introduction to Japanese Linguistics*. Blackwell, 2nd Edition.

UCHIMOTO, K., SEKINE, S. and ISAHARA, H. (2001). The unknown word problem: a morphological analysis of Japanese using maximum entropy aided by a dictionary. In *Empirical Methods in Natural Language Processing (EMNLP)*: 91-99.

UTSURO, T., SHIME, T., TSUCHIYA, M., MATSUYOSHI, S. and SATO S. (2007). Chunking and Dependency Analysis of Japanese Compound Functional Expressions by Machine Learning Text, Speech and Dialogue. In *10th International Conference, TSD* 2007, Plzen.

Wortverbindungsfelder
Fields of multi-word expressions

Annelen Brunner[1], Kathrin Steyer[1]
Institute for the German Language, Mannheim

Abstract

In this paper we outline our corpus-driven approach to detecting, describing and presenting multi-word expressions (MWEs). Our goal is to treat MWEs in a way that gives credit to their flexible nature and their role in language use. The bases of our research are a very large corpus and a statistical method of collocation analysis. The rich empirical data is interpreted linguistically in a structured way which captures the interrelations, patterns and types of variances of MWEs. Several levels of abstraction build on each other: surface patterns, lexical realizations (LRs), MWEs and MWE patterns. Generalizations are made in a controlled way and in adherence to corpus evidence. The results are published online in a hypertext format.

Keywords: multi-word expression, collocation, corpus-driven, usage-based, corpus linguistics, phraseology, lexicology.

1. Methodological approach

We present a structured approach to the study of multi-word expressions (MWEs) which applies a strongly corpus-driven method and results in a novel type of lexicographic description and presentation (*cf.* Steyer and Brunner 2009, Brunner and Steyer 2009).

Based on the concept of *Usuelle Wortverbindungen* (Steyer 2000; Steyer 2004), we regard multi-word expressions as conventionalized patterns of language use that manifest themselves in recurrent syntagmatic structures (*cf.* Feilke 2004). MWEs can comprise fixed lexical components as well as abstract components representing a certain subset of lexical items. Our concept encloses not only idioms and idiosyncratic structures, but all multi-word units which have acquired a distinct function in communication. Real-life usage, pragmatics and context are central to our approach.

In detecting as well as describing these units we work bottom-up in a strongly corpus-driven way (*cf.* Sinclair 1991; Tognini-Bonelli 2001). The following principles, which

[1] {brunner,steyer}@ids-mannheim.de

correspond to the definition of corpus-driven work detailed by Tognini-Bonelli, characterize our approach.

We use the empirical basis of a very large corpus. DeReKo (Deutsches Referenzkorpus, KLa2009), located at the Institute for the German Language (IDS), is the largest collection of written German available today and comprises over 3.7 billion word tokens, mostly from modern newspaper articles. At the current stage we use DeReKo as it is, as our focus is on the model of analysis.

The data is pre-structured by statistical collocation analysis. The algorithm we use ("Kookkurrenzanalyse", Belica 1995) is a sophisticated method which clusters keyword in context (KWIC) lines in several hierarchical levels and also computes the most common order of the surface forms which appear in those clusters (*cf.* KLb2009, Keibel and Belica 2007). The results are a very good basis for our work, as the statistical method shows regularities in the data in a very objective way by considering only word form surfaces. However, we do not take the clusters as they are but use them as a starting point for human interpretation.

Interpreting this rich empirical data we try to take as few pre-conceived notions of how language works as possible and develop the analysis and presentation of the data to fit corpus evidence. We work bottom up from the language surface structure and take monitored steps of interpretation.

In strong adherence to corpus data, we only describe MWEs and variations of MWEs which are attested in our corpus, so the results are always grounded on empirical evidence. As a result of studying corpus data, we came to consider three characteristics as central to the nature of MWEs:

- Usage and context are crucial when identifying and describing MWE entities.
- Most MWEs are variable and can very often be modified and extended in various ways.
- There are rich interrelations between MWEs such as similarities and contrastive nuances in usage, combinations of MWEs which create rich forms of expression and more abstract groups of structurally similar MWEs, which are no longer completely fixed on the lexical surface.

These characteristics are emphasized in our model for describing MWEs.

2. Model of analysis

Our model of analysis has some similarities to that of Hanks detailed in the description of his *Corpus Pattern Analysis* (CPA):

> Concordance lines are grouped into semantically motivated syntagmatic patterns. Associating a 'meaning' with each pattern is a secondary step, carried out in close coordination with the assignment of concordance lines to patterns. The identification of a syntagmatic pattern is not an automatic procedure: it calls for a great deal of lexicographic art. Among the most difficult of all lexicographic decisions is the

> selection of an appropriate level of generalization on the basis of which senses are to be distinguished." (CPA2009)

We, too, have to tackle the task of assigning meaning to syntagmatic patterns and to find the right level of abstraction. CPA aims at describing single words (*cf.* Hanks 2008), while we are interested in MWEs, which adds an additional level of complexity as identifying the surface form itself requires an interpretative effort. To handle the difficulties of generalization, our model has several hierarchical levels which build upon each other. Figure 1 gives an overview of its structure.

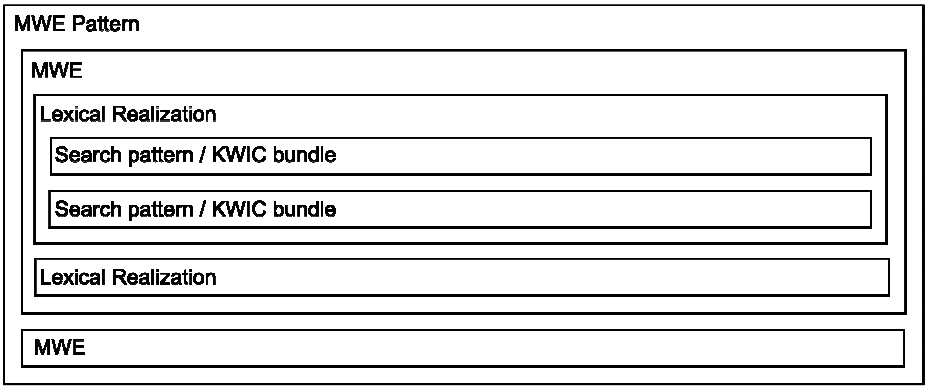


Figure 1. Hierarchical model of analysis

In this paper, we will mostly focus on the example MWE "in den Ohren klingen" [to sound in the ears].

As a starting point, we conduct a collocation analysis of the target word form "Ohren" [ears]. We decided to only use non-lemmatized word form surfaces as targets for the algorithm, as our model of analysis is strongly surface based. The collocation analysis of different inflectional forms of a lemma can result in quite different profiles and we do not want to gloss over these differences too quickly. In this respect, we adhere to Sinclair's claim:

> "There is a good case for arguing that each distinct form is potentially a unique lexical unit, and that forms should only be conflated into lemmas when their environments show a certain amount and type of similarity." (Sinclair 1991: 8)

Collocation analysis outputs several clusters which are relevant for the MWE "in den Ohren klingen", mainly those forming around inflected forms of "klingen" [to sound]. The KWIC lines which comprise these clusters will be the basis of our analysis.

2.1. Search patterns

On the first level, the KWIC lines which have been clustered by collocation analysis are explored and subjected to further structuring. For this task, we use search patterns based on regular expressions.

This step is necessary because collocation analysis shows the relationships between word surfaces, but does not consider the underlying syntactic structure nor can it recognize similarities in meaning and usage. For example, the cluster “Ohren - klingen” [ears - to sound] also contains realizations of other MWEs like “die Ohren klingen” [the ears resound].

At this point, human interpretation builds on the pre-structuring done by statistics. Search patterns can be defined flexibly to capture the structures we are interested in. They serve an analytical purpose, as they allow us to explore possible surface variations, *e.g.* common fillers of slots between fixed elements which can be examined as well.

For example, we find that though the surface form “in den Ohren klingen” [to sound in the ears] is indeed the most common realization of the MWE, the element “den” [the] is quite often replaced in the actual realizations. With the help of the search patterns we explore the fillers for the slot between “in” and “Ohren” and find that three kinds of fillers are most dominant: possessive pronouns, genitive phrases referring to a person and adjectives denoting groups of people, most often referring to their nationalities.

These are example KWIC lines for the three different kinds of surface realizations:

A98/SEP.60063	und aus früheren Tagen **in unsern Ohren klingen**.
P92/FEB.04217	Wie Hohn mußte **in Strolz' Ohren** der Beifall der Tausenden **klingen**
A01/OKT.36053	Die Erklärungen des saudischen Diplomaten mögen **in westlichen Ohren** hohl und feindselig **klingen**

Search patterns allow us to group instances which have similar surface characteristics so that these groups can serve as the basis of further analysis.

2.2. Lexical realizations

Lexical realizations (LRs) are an intermittent step between the hard language surface, as captured by the search patterns and the MWEs. Corpus research clearly shows that the surface form of an MWE is nearly always subject to variation. When generalizing quickly to a single form, many of these nuances are lost. LRs allow us to focus on different typical forms an MWE can take, to show their relationships and to comment on them. An MWE in our model is thus represented by a collection of LRs organized in a tree-like structure.

We distinguish between different kinds of LRs according to a basic set of types which was developed from empirical experience.

For each MWE, a *Core LR* is defined which represents the minimal surface structure necessary to recognize the MWE in its communicative function. Alternative core

realizations can exist, called *Core Variant LRs* in our model. In addition, we define *Extension LRs*, extensions to the core, which can be internal as well as external modifications and additions, *e.g.* prepositional phrases, verbs, modifying adjectives or adverbs. The last type of LR is *Context LR*, defined to highlight word forms which typically appear close to the MWE realization without being part of its structure.

The *LR Group* represents a container which contains all realizations of the MWE. It also serves as a root element of an LR tree. The other LRs can be arranged in several levels.

The LR structure of the MWE "in den Ohren klingen" is shown in Figure 2.

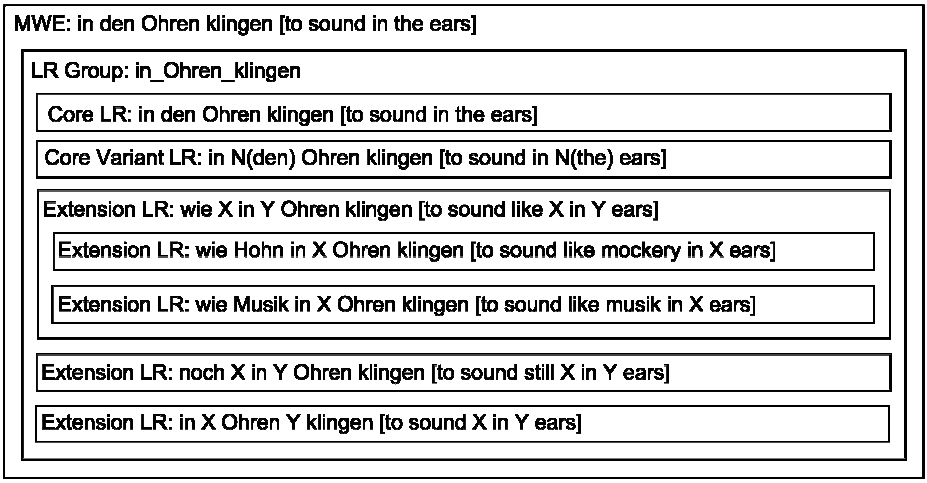


Figure 2. LR tree of the MWE "in den Ohren klingen"[2]

Each LR subsumes the appropriate KWIC lines which are bundled by search patterns and contains more information about the specifics of this realization, for example nuances in meaning and usage and relative frequency information.

LRs can contain slots, which are represented by capital letters in the LR's name. For example, the Extension LR "wie X in Y Ohren klingen" [to sound like X in Y ears] has two slots: Slot Y further specifies the word form "Ohren" [ears] and its fillers are of the kind we found by studying the search patterns in the previous section – most often: "den", possessive pronouns, adjectives.

[2] This example does not contain a Context LR. A typical Context LR would be for example "Knopf im Ohr ... Steiff" [button in the ear … Steiff] which belongs to the MWE "Knopf im Ohr" [button in the ear]. This Context LR highlights a word form, "Steiff" (the name of a toy company), which appears very frequently in the vicinity of the MWE's Core LR. This is an indicator that the MWE is often used to refer to a characteristic of stuffed animals manufactured by the company Steiff, which have a metal button punched into their ear as a brand label.

Slot X is filled by nouns which serve as a simile for how something is received or experienced. Two fillers for this slot are extremely frequent: “Hohn” [mockery] and “Musik” [music]. Because of their typicality the realizations with these fillers are presented as separate LRs, which are dependent on the LR “wie X in Y Ohren klingen”.

Such slots are represented as tables in the LR’s body which list the abstract types and/or concrete lexical items that serve as fillers. These tables are created manually as a result of the study and categorization of the KWIC lines. Only systematic slots, *i.e.* slots with fillers which show some regularity, are represented in this way. They give important insight into the paradigmatic variability of MWEs.

In addition to that, each LR gives direct access to the KWIC lines captured by the search patterns it subsumes and to automatically generated lists of the surface realizations of every underspecified element in these patterns – an unrevised slot-filler list. So it is also possible to take a look at the hard corpus data and see the raw frequencies.

2.3. MWEs, MWE patterns and relationships between them

MWEs are represented by an LR tree as shown in Figure 2 above. In addition, each MWE is assigned a description which contains a paraphrase that is true for all LRs and represents the core meaning of the MWE. For our example “in den Ohren klingen” the general paraphrase would be: “sth is experienced intensely in a certain way and remembered”. Depending on the realization of this rather complex MWE different aspects of this general meaning are emphasized.

In addition to that, an MWE can also contain information about its typical genre, its phrasal structure and its relative frequency in the collocation profile of the target word form.

MWE patterns are an additional step of abstraction which is not obligatory for all MWEs. The patterns are generalizations over structurally similar MWEs and contain at least one underspecified component. Two types of MWE patterns can be distinguished:

1. The MWEs which comprise the pattern are near synonyms and the same meaning can be assigned to all of them. In this case, the meaning paraphrase is assigned to the MWE pattern instead of the separate MWEs.
2. The realizations of the underspecified components are all different in meaning. This results in a group of MWEs which each have a distinct meaning but still have a meaning component in common. The MWE pattern is assigned the most general meaning paraphrase, but each MWE still carries its own meaning paraphrase detailing its specifics.

The example MWE “in den Ohren klingen” can be considered part of an MWE pattern “in den Ohren VERB_Geräusch” [in the ears VERB_sound] and is grouped together

with similar MWEs. These MWEs are not completely identical in meaning – "in den Ohren klingen", which is also the most frequent of the three, has a much richer meaning than the other MWEs. However, in one aspect, they are indeed very similar: They can all express the meaning "sth is experienced intensely (most often acoustically)".

Another important aspect of our model is that interrelations can be defined between MWEs or MWE patterns. These interrelations can be of different kinds, but often involve a similarity in usage or a frequent combination of MWEs or MWE patterns. The interrelation structure of "in den Ohren klingen" is represented in Figure 3.

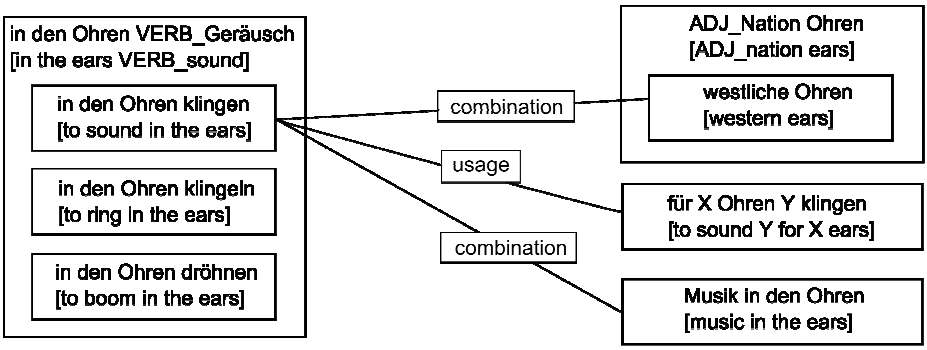


Figure 3. Interrelations between MWE "in den Ohren klingen" and other MWEs and MWE patterns

"In den Ohren klingen" is often combined with "Musik in den Ohren", resulting in the form "wie Musik in den Ohren klingen" [to sound like music in the ears]. With the MWE pattern "ADJ_Nation Ohren" it is combined to form the realization "in ADJ_Nation Ohren klingen" [to sound in ADJ_nation ears]. The MWE "für X Ohren Y klingen" [to sound Y of X ears] is very similar to one meaning aspect of "in den Ohren klingen": "to be experienced in a certain way by a certain group of people".

3. Implementation and presentation

For our analysis, we use a specially developed software tool, which takes collocation clusters as input and is used to match, group and annotate the KWIC lines according to the model described above. The analyzed data are stored in an XML format which allows different modes of visualization.

Currently, our results are presented as fields of MWEs ("Wortverbindungsfelder"), each centered on a specific word form. The hierarchical structures and interrelations between the different units are realized in a hypertext format and direct access to structured corpus data is provided. All levels of description are enriched by lexicographic comments like the description of meaning and usage in the corpus. Thus

the results can be viewed in two different ways. On the one hand, the structure allows for the reconstruction of the typical usage of MWEs from the corpus data and provides a complete documentation of our interpretative method. On the other hand, the narrative comments allow an access more similar to that of traditional lexicographical products. The first version of fields of MWEs, one centred on forms of the word "Grund" [ground/reason] and one centered on forms of the word "Ohr" [ear] are available on the internet, accessible from our site "Wortverbindungen online": http://wvonline.ids-mannheim.de/

Though developed in an experimental research context, we believe that our approach can give valuable impulses to lexicographic practice: Working with real-life data helps revising common misapprehensions about the structure and meaning of MWEs and results in a new form of presentation, highlighting the importance of variability, context and usage. In addition to that, our model presents a novel approach in including corpus data not only as illustration, but as a basis of description, and offers structured access to real-life data, taking advantage of the options of the electronic hypertext format.

References[3]

BELICA, C. (1995). *Statistische Kollokationsanalyse und Clustering. Korpuslinguistische Analysemethode*. Mannheim: Institut für Deutsche Sprache: http://www.ids-mannheim.de/kl/projekte/methoden/ur.html.

BRUNNER, A. and STEYER, K. (2009). A Model for Corpus-Driven Exploration and Presentation of Multi-Word Expressions. In J. Levická and R. Garabík (eds). *NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference Smolenice, Slovakia, 25-27 November 2009. Proceedings.* Bratislava: Tribun: 54-64.

CPA2009. *Corpus Pattern analysis.* Internet: http://nlp.fi.muni.cz/projekty/cpa/.

FEILKE, H. (2004). Kontext – Zeichen – Kompetenz. Wortverbindungen unter sprachtheoretischem Aspekt. In K. Steyer (ed.). *Wortverbindungen – mehr oder weniger fest. Jahrbuch des Instituts für Deutsche Sprache 2003*. Berlin/New York.

HANKS, P. (2008). Lexical Patterns. From Hornby to Hunston and Beyond. In E. Bernal and J. DeCesaris (eds). *Proceedings of the XIII Euralex International Congress, Barcelona, 15-19 July 2008.* Barcelona: Institute for Applied Linguistics, Pompeu Fabra University: 89-129.

KEIBEL, H. and BELICA, C. (2007). CCDB: A Corpus-Linguistic Research and Development Workbench. In *Proceedings of Corpus Linguistics 2007*, Birmingham. http://corpus.bham.ac.uk/corplingproceedings07/paper/134_Paper.pdf.

KLA2009. *Ausbau und Pflege der Korpora geschriebener Gegenwartssprache. Das Deutsche Referenzkorpus – DeReKo.* Mannheim: Institut für Deutsche Sprache: http://www.ids-mannheim.de/kl/projekte/korpora/.

KLB2009: *Methoden der Korpusanalyse- und erschließung.* Mannheim: Institut für Deutsche Sprache: http://www.ids-mannheim.de/kl/projekte/methoden/.

SINCLAIR, J. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

[3] All hyperlinks checked on 14 December 2009.

STEYER, K. (2000). Usuelle Wortverbindungen des Deutschen. Linguistisches Konzept und lexikografische Möglichkeiten. *Deutsche Sprache*, 28(2): 101-125.

STEYER, K. (2004). Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikografische Perspektiven. In K. Steyer (ed.). *Wortverbindungen – mehr oder weniger fest. Jahrbuch des Instituts für Deutsche Sprache 2003*. Berlin/New York: DeGruyter: 87-116.

STEYER, K. and BRUNNER, A. (2009). *Das UWV-Analysemodell. Eine korpusgesteuerte Methode zur linguistischen Systematisierung von Wortverbindungen*. Mannheim: Institut für Deutsche Sprache: http://www.ids-mannheim.de/pub/laufend/opal/privat/opal09-1.html.

TOGNINI-BONELLI, E. (2001). *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins Publishing Company (*Studies in Corpus Linguistics* 6).

SSLD
a French SMS to Standard Language Dictionary

Louise-Amélie Cougnon[1], Richard Beaufort[1]
Université catholique de Louvain - CENTAL

Abstract

This paper presents a methodology to semi-automatically build up a dictionary out of an SMS corpus. First, we describe the three-step approach that extracts the dictionary entries from the corpus and we detail the smart manual sorting performed on the dictionary. Then, we give a panorama of SMS phenomena observed in the dictionary. Finally, we survey the current limits of our methodology and the related improvements that can be made to it.

Keywords: SMS dictionary, SMS phenomena, alignment, finite-state machines.

1. Introduction

At a time when technology intensifies and strengthens mechanical and human communication over the world, the study of new types of dictionaries and lexical resources seems essential. The language used in Short Message Service (SMS), on the same level as *chat* language, is one of these new written forms of communication. When dealing with SMS[2], one has to cope with various issues: new linguistic phenomena, language processing difficulties and lexical resource limits. Linguistic phenomena in SMS go from phonetic and numeral scripts, abbreviations and capital letters, to intensive use of neologisms, language mixing and borrowing, through new code systems such as emoticons. Processing SMS corpora involves identifying lexemes, applying dictionaries and using particular tools such as taggers, grammatical analyzers and lexical resources. There is a wide range of lexical resources for SMS studies, but unfortunately, studies on transcription from SMS to standard language are few and results are still too basic (mainly because they are based on corpora of limited size).[3]

The *sms4science* project aims at collecting international SMS corpora. Since the beginning of the project, we have been questioning the usefulness of SMS corpora and

[1] {louise-amelie.cougnon,richard.beaufort}@uclouvain.be

[2] The acronym refers to the service as much as to the messages exchanged during the service.

[3] Guimier de Neef and Fessard (2007) is a notable exception as they made use of a corpus of 10,000 SMS.

SMS to standard language transcription. At this stage, we had already worked on SMS transcription, especially on the reverse dictionary, viz. standard to SMS language (http://www.uclouvain.be/44009.html). Then, in early 2008, a new project was set up within the framework of our research centre: *Vocalise*, an SMS-to-Speech synthesis project (http://cental.fltr.ucl.ac.be/projects/vocalise/index_EN.html). In order to improve SMS speech synthesis, the new project developed an SMS word alignment system based on a corpus of 30,000 text messages and their manual transcription.[4] It was, for us, the opportunity to address the question of SMS to standard language transcription again. We decided to use the *Vocalise* aligned corpora to draw up an SMS to Standard Language Dictionary (SSLD). In order to meet this objective, we built a list of entries based on all the words of the aligned corpora.

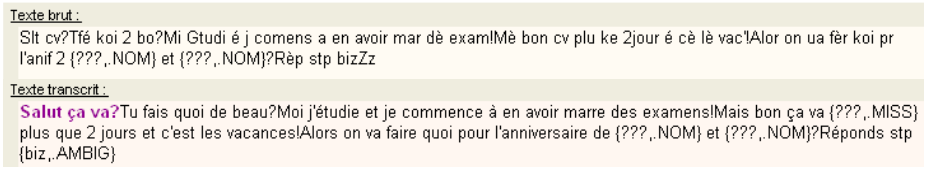
This paper is organized as follows. Section 2 presents the three-step approach, which made it possible to semi-automatically build up a dictionary out of an SMS corpus, while Section 3 focuses on the smart manual sorting of the dictionary and presents the SMS phenomena (which we refer to as "categories") that constitute the dictionary entries. Section 4 details the kinds of mistakes our three-step approach made at different levels, and proposes some possible improvements, which should significantly enhance the SSLD-making procedure. We finally draw some conclusions in Section 5.

2. From SMS-gathering to dictionary-making

Three distinct steps enabled the dictionary making: corpus collection and transcription (part of the sms4science project), corpus alignment (part of the Vocalise project) and raw SMS resource extraction.

2.1. Corpus collection and transcription

We built up the SSL dictionary from a French SMS corpus of 30,000 messages, gathered in Belgium, semi-automatically anonymized and manually normalized[5] at the Université catholique de Louvain (Fairon and Paumier 2006). As shown in Figure 1, the SMS corpus and its transcription constitute parallel corpora aligned at the message-level.


Texte brut :

Slt cv?Tfé koi 2 bo?Mi Gtudi é j comens a en avoir mar dè exam!Mè bon cv plu ke 2jour é cè lè vac'!Alor on ua fèr koi pr l'anif 2 {???,.NOM} et {???,.NOM}?Rèp stp bizZz

Texte transcrit :

Salut ça va?Tu fais quoi de beau?Moi j'étudie et je commence à en avoir marre des examens!Mais bon ça va {???,.MISS} plus que 2 jours et c'est les vacances!Alors on va faire quoi pour l'anniversaire de {???,.NOM} et {???,.NOM}?Réponds stp {biz,.AMBIG}

Figure 1. Snapshot of message-level aligned corpora

[4] The project "Faites don de vos SMS à la science" collected 30,000 French text messages in 2004.

[5] *"SMS normalization consists in rewriting an SMS text using a more conventional spelling, in order to make it more readable for a human or for a machine"* (Yvon 2008).

2.2. Corpus alignment

Unfortunately, this message-level alignment does not allow for pertinent lexical extraction and equivalence. In order to achieve this purpose, we needed an alignment at the word level: for each word of a sentence in the standard transcription, we had to know the corresponding sequence of characters in the SMS version. As an accurate automatic linguistic analysis of the SMS corpus was not possible, we needed another way of producing this word-alignment: a method able to align sentences at the character level. This method is called "string alignment".[6] One way of implementing this string alignment is to compute the edit-distance of two strings, which measures the minimum number of operations (substitutions, insertions, deletions) required to transform one string into the other (Levenshtein 1966). Using this algorithm, in which each operation gets a cost of 1, two strings may be aligned in different ways with the same global cost. For instance, the couple (*kozer*, *causé*) could be aligned:

```
(1) ko_ser      (2) k_oser      (3) ko_ser      (4) k_oser
    causé_          causé_          caus_é          caus_é
```

where underscores (_) mean "insertion" in the upper string, and "deletion" in the lower string. However, from a linguistic standpoint, only alignment (1) is desirable, because corresponding graphemes are aligned on their first character. In order to automatically choose this preferred alignment, we had to distinguish the three edit-operations, according to the characters to be aligned. For that purpose, probabilities were required. Computing probabilities for each operation according to the characters to be aligned was performed through the following iterative algorithm, implemented in the framework of the Vocalise project:

STEP 1. Align the corpora using the standard edit-distance (with edit-cost of 1).

STEP 2. From the alignment, learn probabilities of applying a given operation on a given character.

STEP 3. Re-align the corpora using a weighted edit-distance, where the cost of 1 is replaced by the probabilities learned in STEP 2.

STEP 4. If two successive alignments provided the same result, there is a convergence and the algorithm ends. Else, it goes back to STEP 2.

Hence, the algorithm gradually learns the best way of aligning strings. On our SMS parallel corpora, the algorithm converged after seven iterations and provided us with a result (see Figure 2) from which the lexicon of SMS words could be built.

A standard way of implementing edit-distance is to use dynamic programming (Viterbi 1967). However, in order to easily compute weighted edit-distances, we used weighted finite-state machines, which were shown by Mohri (2003) to be very efficient in this

[6] String alignment comes from bioinformatics, where sequences of DNA must be arranged in such a way that similarities and differences are identifiable.

task. The finite-state library in use here is described in Beaufort (2008), and the finite-state alignment of the iterative algorithm (steps 1 and 3) is detailed in Beaufort *et al.* (2008).

```
28620: S_t_t c__v_?_T__fé__ k_oi 2_ b_o_?_M_i G__tudi_ é_ j_ com_ens_ a
28620: Salut ça va? Tu fais quoi de beau? Moi j'étudie et je commence à
```

Figure 2. Snapshot of the word-level alignment corpora, where symbols _ stand for insertions and deletions

2.3. SSLD input extraction

Based on this character-level alignment, an extraction script[7] enabled us to extract, for each sequence, its raw and standard variants. The script loaded a regular French language dictionary[8] that allowed matching our SMS standard sequences with recognised inflected forms and their lemma. In our SSLD, each entry is not followed by its standard sequence, but by its lemma, as can be seen in Figure 3.

```
monitric | ~~(monitrice)~~ | moniteur | N+z1:fs
```

Figure 3. SSLD extract showing the unwanted (standardised inflected) column

For ambiguous sequences that showed various lemmas, a new entry was created for each possible grammatical interpretation. Figure 3 also shows that the SMS sequences and the lemma are followed by their grammatical and inflectional information and potentially, by additional information, such as lexical layers (z1, z2, z3, etc.) and semantic information (as *Hum* for any name referring to a person or *Profession* for any name referring to a profession, etc.).[9]

The extraction script mainly implements the following algorithm:

```
STEP 1. For each aligned pair {SMS message, standard message},
        Split the two messages according to blanks and punctuations
         in the standard message
        For each pair of {SMS, standard} segments
        Clean segments (remove insertion and deletion symbols _,
         convert each upper case into the corresponding lower case)
        Store the pair in a temporary lexicon, except if the SMS
         sequence is empty or matches with a number/time pattern
STEP 2. For each stored pair from the temporary lexicon,
        If the standard word exists in the DELAF lexicon,
         for each DELAF lexicon entry {standard word, lemma,
         category}, create a new SSLD entry {SMS sequence, lemma,
         category}
        Else, create a new SSLD entry, {SMS sequence, UNKNOWN tag}
```

[7] Our gratitude goes to Hubert Naets, who wrote this script.

[8] The DELAF was our reference dictionary; it is an electronic dictionary for French, initially built up by M. Gross and mainly developed during the 80's and 90's. It includes 683,824 entries for 102,073 different lemmas.

[9] Our system of codes is totally inspired by Unitex dictionaries syntax.

After the application of this script on our aligned corpora, the SSLD lexicon comprised 45,049 entries for 10,318 different lemmas.[10]

3. Smart sorting and analysis of the SSLD

3.1. Smart sorting

At this step, we manually filtered out unwanted entries so as to obtain a smarter SSLD. All unknown sequences added to the SSLD by the extraction script were manually revised: neologisms (later than 2001)[11], word plays, proper names (toponyms, first names and trade marks), foreign words (*monkey*, *besos*, *aanwezig*, etc.), unrecognised sign/number patterns (*e.g. 07h5* for *07h05*), emotive graphics (*e.g.* repetition of letters showing intensity) and transcriber's mistakes (*cnpine* for *copine* 'girl friend', *premdr* for *prendre* 'to take', etc.). All these categories were kept, apart from proper names and transcriber's mistakes.

During this checking task, each SSLD entry was also labelled with one of the seven SMS categories (presented in section 3.2) we defined in order to characterize the stylistic phenomena of the SMS corpus. Some ambiguous sequences, however, could not be directly associated with any of our categories, and we had to go back to the initial corpus and look at the context. For instance, the entry *rè*, whose lemma was *trait*, was difficult to label: we clearly could have thought of an abbreviating phenomenon (added to some sort of phonetisation), while *rè* was just the last segment of the SMS form *pRmeterè*, which stood for *permettrait* ('would permit') and had been wrongly segmented into 2 entries by the extraction script.

3.2. Analysis

3.2.1. Seven SMS categories

First of all, and contrary to what one might think, standard inflected words that satisfy standard spelling make up half of our SSLD entries. On the other half of the entries, some SMS phenomena were rapidly recognized: the abbreviating process is commonly known, as well as phonetisation (which is a subcategory of abbreviation), which describes letters, numbers or signs used for their phonetic values.[12] We chose to distinguish the use of signs and the use of numbers. We finally added the "mistakes" category (which includes SMS user, transcriber, word-aligner or algorithm mistakes) and the "unlikelies" category, which are not SMS phenomena strictly speaking but which have to be considered apart from other SMS phenomena. None of these

[10] Our gratitude goes to Master students in philology who helped us sorting out the dictionary entries.

[11] Unfortunately, the DELAF dictionary has not been significantly upgraded since 2001.

[12] A letter used for its phonetic value is spelt, instead of being simply pronounced like in word context.

categories was deleted as they all conveyed specific information that could be used to improve automatic SMS reading and understanding.

The phonetisation category had to be specified. Since we decided to put numbers and signs aside, this category was used to define any sequence that phonetically resembled the standard word. We put in this category phonetisation strictly speaking (*e.g. pnible* for *pénible*), any sequence showing schwa deletion (*e.g. bêtis* for *bêtise*), but also any simplification that maintains the phonetic resemblance (*e.g. ail* for *aille*, the subjunctive of *aller*, "to go"). This category is by far the most popular SMS graphic phenomenon, because it includes any unaccentuated word.

3.2.2. The "unlikelies"

The fact that, for ambiguous terms, a new entry is created for each possible lemma, ensures a certain improvement of the dictionary; but it also adds some ambiguity if, for example, the SSLD was to be used for automatic translation. For terms which could be either nouns or inflected verbs (*e.g. échange*), the ambiguity has to be maintained and could probably be solved by the context. But in other cases, the confusion is unnecessary, because one of the lemmas is very frequent, while the others are fairly rare, at least in SMS context. This is what we called an "unlikely": a rare lemma. All unlikelies were deleted from the dictionary.

Example	**Meaning**
`ballons,baller.V:P1p:Y1p`	"to dance, to jolt"
`muchas,mucher.V:J2s`	"to hide"

Figure 4. Examples of unlikelies in the SSLD

The second example of Figure 4 is of a particular interest: the French homograph of this Spanish word is not frequent enough to maintain an entry in the SSLD dictionary. Nevertheless, we decided not to delete this kind of entries, but to mark them with a special *unlikely* tag that would allow us to identify and delete them later.

3.2.3. Unknown sequences

As we reported above, a sizeable part of unknown words that we reintroduced in the dictionary were words that entered the French language after 2001. These words mostly refer to new realities (*fitness*, *monoparentalité*), or technologies (*adsl*, *bipeur*[13], *pirater*). Some of them, however, are just new labels for well-known realities (*criser* "to be on edge", *tilter* "to suddenly understand", *cafariser* "to sadden", or *moisversaire* "a celebration that happens the same day of each month").

[13] The verb *biper* can be found in the DELAF but not the noun *biper* and its alternative spelling *bipeur*.

Some other sequences labelled as unknown turned out to belong to some specific terminology: *acerifolia* (botany), *markopsien* (marketing) and *émollience* (cosmetics) are good examples of this phenomenon. We decided to keep them as part of the SMS user's lexicon.

Finally, a lot of unknown entries were identified as regionalisms, and included in our final dictionary. As our corpus was collected in Belgium, regionalisms were mostly Belgian or at least shared by Belgium and other French-speaking areas. Words like *baraki*, *berdeller*, *copion*, *guindaille* and *se faire carotter* illustrate this clear trend.

4. Issues and possible improvements

This section presents the different kinds of mistakes that occurred at various stages of our methodology, and the possible solutions we propose to solve them.

4.1. Manual transcription

As a matter of fact, first mistakes are due to the transcriber himself. Even when he carefully checks his work, a single transcriber is not enough to avoid accidental mistakes, which of course occurred quite frequently for a 30,000 SMS corpus. Naturally, we could help the transcriber by checking his transcription several times. However, to err is human, and even multiple checking will not point out all mistakes. A complementary solution could be to automatically perform lexicon looks-up during the transcription process, and to draw the transcriber's attention to possible out-of-vocabulary words or infrequent forms.

4.2. Alignment algorithm

Three kinds of mistakes are due to the alignment algorithm. First, cases of agglutination are frequent: the aligner shows a clear tendency to align on the first of two words when a letter is missing (*cf.* Figure 5). Second, some typography is not handled, such as the *&* symbol not recognized as *et*, or the digit *1* identified as being the letter *i* (*cf.* Figure 6). Third, some subtle cases of phonetisation are not taken into account by the process. This is the case with letters, numbers or signs that replace more than one word.

```
D__t__t__Facon_J_en_Ai Plu__Besoin:-D D__c_Fo___PluS__tréssé_..
De toute façon j'en ai plus besoin:-D Donc faut plus stresser..
```

Figure 5. False alignments

```
G____besoin_2__partaG__ k___k1_s____tan__ a__c toi
J'ai besoin de partager quelques instants avec toi
```

Figure 6. Typography problems

These errors are due to the fact that the alignment works without resort to linguistics; it simply iteratively computes affinities of association between *letters*, and uses them to gradually improve the character-level alignment. However, as recent linguistic studies showed, phonetic transcriptions (*sré* instead of *serai*, '[I] will be', *kom* instead of *comme*, 'as') and phonetic plays (*2m1* instead of *demain*, 'tomorrow', *k7* instead of *cassette*, 'tape') are very frequent in SMS. This could be exploited by the alignment, which could perform its task *through* a phonetic version of the sequences to be aligned. Figure 7 gives an example of phonetic alignment that solves a kind of error depicted in Figure 6.

```
      SMS text:                k _ _ _ k _ _ _  _  1 _ s t a n _ _
      SMS phonetisation:       k _ _ _ k _ _ _  _  e~_ s t a~_ _ _
      Standard phonetisation:  k _ E l k _ @ z  _  e~_ s t a~_ _ _
      Standard text:           q u e l q u e s ' ' i n s t a n t s
```

Figure 7. Phonetic alignment. The phonetic alphabet in use is SAMPA.

Of course, here, an important fact must be taken into account: while a standard written sentence can be automatically analyzed and unambiguously phonetised by NLP applications, it is not the case for an SMS sentence, which is difficult to analyze, and should thus be transcribed as a lattice of possible phonetisations. The alignment, here, will thus face another problem: the weight of these concurrent phonetisations, in order to choose the best path in all possible phonetic alignments.

4.3. Extraction algorithm

The extraction algorithm also showed some limits. First issues are due to the deletion of characters considered as separators: some ambiguous characters considered as separators were lost, while they were used as signs for phonetic purposes or abbreviation (*cf.* Figure 8). However, keeping extra punctuation would have generated too much noise.

```
      Ben viens-chercher-la' clé usb au -sud. 18-à tout de suite.
      Ben viens-chercher-la_ clé USB au _Sud_ 18-À tout de suite.
>>
      -sud,sud.N+z1:ms
      -sud,sud.A+z1:ms:fs:mp:fp
```

Figure 8. Punctuation mismatch

The second loss of information is due to the systematic neutralization of the case, as most upper-case characters were at the beginning of sentences. Nevertheless, some upper case letters carried pieces of phonetic information that would have been useful in the reading of dictionary entries (*e.g.* the *T* in *arT* for *arrête* is always upper case).

The third problem related to identical buffers void of letters or numbers. While it was needed to delete any number or time expression from our dictionary, it was also

unfortunate to lose all character sequences that could have carried information (*e.g.* emoticons).

Actually, all these limitations have a single origin: the extraction algorithm rates a couple of aligned sentences just as two strings of characters, and makes arbitrary choices only based on predefined sets of characters (letters, punctuations, symbols, etc.), without taking the context into account. Based on this observation, we consider the possibility to provide the algorithm with an automatic morphosyntactic analysis of the normalized side of the alignment. This linguistic analysis should help the algorithm split the sentence into the right segments, and add the right entries to the SSLD.

4.4. False entries

Plays on letters were hardly dealt with by the system, because even when both the alignment and the extraction steps did not generate errors, some sequences did not correspond to lexical entries and should have been left out of the dictionary (*cf.* Figure 9). Just as the extraction algorithm, false entries could be rejected by the system, by checking their linguistic analysis through an automatic analyzer.

```
7______rop b_o_ 7____ idylle k_i 7__ternise
C'est trop beau cette idylle qui s'éternise
```

Figure 9. False entries

5. Conclusions and prospects

The dictionary built up using this framework is not exhaustive at all: it covers neither all lexical fields, nor the whole lexicon of any particular field. However, it gets credit for covering an important part of Belgian SMS users' lexicon. It might thus be of interest to have it further examined and compared to standard dictionaries. Which proportion of a standard dictionary is really covered? Are there new spellings and new words in this dictionary – and not in standard ones – which should be included in them? Sequences like *asap*, *lol*, *mdr* (*mort de rire*, stands for *lol* in French), *admin*, are commonly understandable and are not local, or part of a specific terminology or register. Could a standard dictionary be inspired by our SSLD which is based on genuine written practices?

Some improvements of our dictionary might be beneficial. At first, our methodology will be improved, starting with a special focus on the problems raised in Section 4. But we will also enhance our lexicon, by applying our methodology to three new significant French-speaking corpora, gathered in Québec (2009), Switzerland (2009) and France (2010). Finally, we will improve our results by applying a more recent electronic dictionary. In order to improve the dictionary, further studies could also focus on the use of cases (are capital letters always phonetisations?) and specific

characters (is the absence of schwa, like in the opposition *échang* – noun – and *échange* – verb –, significant for grammatical desambiguation?).

Finally, the SSLD is not only a starting point for linguistic studies. This resource is also fundamental for SMS-based applications, like text-to-speech synthesis applied to SMS messages. The automatic linguistic analyzer included in any speech synthesiser uses lexical resources to both disambiguate and phonetise the words that must be read aloud by the system. Faced to SMS messages with noisy forms, an analysis that only relies on standard lexica will fail, while a specialized dictionary like the SSLD should make it easier to find out the standard written word hidden behind a given noisy form. In this context, a reliable SSLD should thus be a real improvement. The Vocalise project, which provided us with the alignment algorithm, is based on this assumption.

References

BEAUFORT, R. (2008). *Application des machines à états finis en synthèse de la parole. Sélection d'unités non uniformes et correction orthographique*. PhD Thesis. Department of Semantics and Computational Logic, Faculty of Computer Science, FUNDP, Namur, March 4th 2008.

BEAUFORT, R., ROEKHAUT, S. and FAIRON, C. (2008). Définition d'un système d'alignement SMS/français standard à l'aide d'un filtre de composition. In *Proceedings of the 9th International Conference on the Statistical Analysis of Textual Data (JADT 2008)*. Lyon. March 12-14, 2008.

COUGNON, L.-A. (forthcoming). La néologie dans 'l'écrit spontané'. Etude d'un corpus de SMS en Belgique francophone. In *Actes du Congrès International de la néologie dans les langues romanes*. Barcelone. 7-10 mai 2008.

FAIRON, C., KLEIN J.R. and PAUMIER S. (2006a). *Le langage SMS*. Louvain-la-Neuve, Presses universitaires de Louvain (*Cahiers du Cental*, 3.1).

GUIMIER DE NEEF, E. and FESSARD, S. (2007). Évaluation d'un système de transcription de SMS. In *Actes du 26e Colloque international Lexique Grammaire*, Bonifacio, 2-6 octobre 2007: 217-224.

LEVENSHTEIN, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics*, 10: 707-710.

MOHRI, M. (2003). Edit-distance of weighted automata: General definitions and algorithms. *International Journal of Foundations of Computer Science*, 14(6): 957-982.

PANCKHURST R. (2008). Short Message Service (SMS) : typologie et problématiques futures. Arnavielle T. (coord.), *Polyphonies*, pour Michelle Lanvin, Montpellier, Éditions LU: 33-52.

VITERBI, A.-J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2): 260-269.

Extracting a bilingual transdisciplinary scientific lexicon

Patrick Drouin[1]
Université de Montréal, Montréal

Abstract

In this paper, we propose a first step leading to the description of the lexicon of scientific language by extracting a transdisciplinary scientific lexicon (TSL) in French and English. We consider that the TSL is domain independent and forms a central lexical core of all domains; it is at the center of the argumentation, the description and the structuring processes we observe in scientific discourse. In order to extract the transdisciplinary lexicon, we use natural language processing tools and statistical techniques. Our identification method relies on two main ideas, the distribution of the TSL in all scientific documents and the specificity of the TSL to scientific documents.

Keywords: lexicography, lexicon, scientific discourse, natural language processing.

1. Introduction

Most linguistic studies dealing with specialized language are interested in subject specific lexicon or terminology, which leads to a general lack of description of other types of lexicon contained in specialized corpora. The exception is the work being done in the area of specialized language teaching like the studies of Coxhead (1998, 2000); but in most cases, as pointed out by Tutin (2007), the lexicon itself is not what is being studied.

In this paper, we propose a first step leading to the description of the lexicon of scientific language by identifying a transdisciplinary scientific lexicon (TSL) in French and English. We consider the TSL to be domain independent and to form a central lexical core of all domains; thus being at the center of the argumentation, the description and the structuring processes we observe in scientific discourse. In order to gather this transdisciplinary lexicon, we use natural language processing tools and statistical techniques. Our identification method relies on two main concepts: **specificity** and **distribution**.

The main idea behind specificity is that we believe that the lexicon we want to describe is highly characteristic of scientific documents and that it can be isolated

[1] Observatoire de linguistique Sens-Texte, patrick.drouin@umontreal.ca

using statistical techniques based on word frequencies. One could easily argue that the most specific lexicon of a scientific corpus will be built around the terminology of all domains represented in the corpus. We tend to agree with the argument and this is the reason why we combine specificity and distribution, the latter being the proportion of sub-corpora in which a word appears based on the total number of corpora. Once again, this clue can be observed in corpora using simple frequency measures.

2. Methodology

Evaluating the specificity of a word for a corpus requires a point of comparison. For the current study, we will be using, as a complement of our scientific corpora, newspaper corpora as a reference point. All corpora are described in Section 2.1. Sections 2.2 and 2.3 explain how specificity and distribution are measured.

2.1. Corpora

2.1.1. Scientific corpora

Our study is based on specialized corpora built from PhD theses and scientific papers; thus, it is at the moment somehow genre specific. Our corpora are open corpora and we will include more genres as time passes. So far, the nine subject areas covered by the corpora are: archaeology, chemistry, geography, history, computer science, engineering, law, physics and psychology.

We rely on comparable corpora totaling approximately 4 million words in French and English. All subject areas are evenly represented with roughly 200,000 words from thesis and 200,000 words from scientific papers. Although the data is not exactly the same in both languages, we were able to establish a balanced bilingual corpus. All documents included in the corpora were published between 1997 and 2007 since older documents are harder to find in electronic form. We believe that the 10-year time span is narrow enough to proceed to a synchronic study as the one proposed.

Domain	Word Count
Archeaology	233699
Chemistry	213239
Computer Sciences	207445
Engineering	238868
Geography	227715
History	245014
Law	234784
Physics	214546
Psychology	245292
Total	2060602

Domain	Word Count
Archeaology	254956
Chemistry	191034
Computer Sciences	247612
Engineering	145252
Geography	220653
History	320267
Law	374830
Physics	197867
Psychology	360473
Total	2312944

Table 1. French Corpus (Papers / Theses)

Tables 1 and 2 give further details about the scientific French and English corpora we build. The left side contains information about the scientific papers while the left side describes the theses. The total number of words in the French corpus is 4,373,546 while the English corpus is slightly smaller at 4,066,759 words.

Domain	Word Count
Archeaology	226470
Chemistry	206616
Computer Science	210649
Engineering	224504
Geography	227887
History	222889
Law	238867
Physics	215145
Psychology	242847
Total	2015874

Domain	Word Count
Archeaology	259264
Chemistry	224647
Computer science	238250
Engineering	199606
Geography	245391
History	222241
Law	242857
Physics	196559
Psychology	222070
Total	2050885

Table 2. English Corpus (Papers / Theses)

The preprocessing of documents of all corpora was performed using freely available tools. The first step in preparing the documents was handled by the TreeTagger (Schmid 1994), a part-of-speech (POS) tagger. In order to be able to establish the real frequency of words contained in our corpora, we decided to simplify the tagging done by TreeTagger and to simply keep the lemma and the POS tag thus discarding the actual inflected form from the corpus. Using such a simplification allows us to compute frequencies of lemmas instead of multiple inflected forms for the same word.

2.1.2. Reference corpora

The bilingual aspect of our study implies that we have a reference corpus for each language. The French reference corpus is built from about 12 million words taken from articles published in 2002 in the newspaper *Le Monde*. As far as English is concerned, we used parts of the *British National Corpus* (BNC) in order to come up with a corpus comparable to the one we used in French. We divided the BNC into genres using David Lee's classification (2001). This researcher has divided the BNC corpus into 46 genres and 8 super genres:

- academic prose;
- non-printed essays;
- fiction;
- letters;
- broadsheet national newspapers;
- regional and local newspapers;
- tabloid newspapers;
- non-academic prose (non-fiction).

Since we wanted to build reference corpora of decent proportions, we only considered super genres that contained over twelve million words:

- academic prose, which contains more than 12 million words;
- fiction, which contains more than 18 million words;
- non-academic prose, which contains more than 19 million words.

For our experiment, we decided to use only the texts that belonged to super genres *broadsheet national newspapers*, *regional and local newspapers* and *non-academic prose (non-fiction)* as they allowed us to gather a reference corpus that was quite similar to *Le Monde* both in size and in genre. All corpora were truncated to exactly twelve million words.

2.2. Specificity Testing

Corpus specificity is evaluated using a statistical measure proposed by Lafon (1980) called *specificity test* (*calcul des spécificités*). This measure allows us to compare the frequency of a word in a corpus (here our scientific corpora or SC) to the frequency of the same word in another corpus (our reference corpora or RC).

Corpus	**RC**	**SC**	**Total**
Frequency of word	a	b	a+b
Frequency of other words	c	d	c+d
Total	a+c	b+d	N=a+b+c+d

Table 3. Contingency table used to describe frequencies in corpora

The actual calculation is performed using the following formula (Lafon 1980):

$$log\ P(X=b) = log\ (a+b)! + log\ (N-(a+b))! + log\ (b+d)! + log\ (N-(b+d))! - log\ N! - log\ b! - log\ ((a+b)-b)! - log\ ((b+d)-b)! - log\ (N-(a+b)-(b+d)+b)!$$

This technique pinpoints three types of words based on their frequency: positive, negative or neutral specificities. The first ones have a frequency which is higher than could be expected in the SC based on a normal distribution evaluated from observations made in the RC, they are thus highly specific. The second ones have a frequency that is lower than expected and are thus significantly absent from the SC. The last ones have a frequency in the SC that is in normal range.

2.3. Distribution Testing

There are at least two ways to look at distribution when dealing with corpora like the ones we use for our experiment. The first approach would be to create sub-corpora based on subject matter and to look at the relative frequency of the words in these sub-corpora, which might have different sizes but are homogenous from a content

perspective. The other method would consist of dividing the whole scientific corpora in chunks (heterogeneous content), making sure that sub-corpora would have the same size (computed in words) and then comparing the frequencies in the different parts of the corpus. In this way, we can compare raw frequencies of words across the corpus without taking into account the size of the sub-corpora. For the current experiment, we decided to use the *natural* division of the scientific corpora and use the various domain specific sub-corpora as our unit of measure. Although the size of sub-corpora is variable, it remains comparable. The variation is taken into account by the statistical test.

Since words can be highly specific to our specialized corpora but still be linked directly to one of the 9 subject areas (in other words, they can be terms), we want to make sure that words retained as potential TSL units are distributed in our SC. In order to be included in our list, a word both needs to appear in all subject areas of the specialized sub-corpora and to have a high-specificity level.

3. Results and discussion

3.1. Specificity

Since lexical items that are highly characteristic of the scientific corpora interest us, we will focus solely on positive specificity for the description of our TSL. Table 4 contains, in decreasing order of specificity, the top 25 words retrieved from our English scientific corpus. Although a few of the forms listed could be highly polysemic such as *system* or *process*, we can safely say that most data retrieved by the specificity test could be included in a TSL.

Word	Arch.	Chem.	Law	Geo.	Hist.	Comp. Sci.	Eng.	Phy.	Psy.
model	66	567	230	244	39	786	758	711	623
analysis	339	337	155	392	52	373	419	246	384
function	66	252	73	70	33	324	420	346	489
phase	83	419	40	31	8	127	227	785	291
system	183	489	1454	316	266	742	1132	1121	247
structure	126	469	92	183	59	277	411	341	117
method	59	303	58	80	49	380	377	175	210
state	122	308	912	360	417	379	119	1036	62
design	25	127	109	71	20	364	1016	288	156
interaction	36	173	17	148	13	126	218	253	213
research	378	107	769	498	21	327	283	46	476
surface	205	656	8	131	9	29	462	279	22
order	169	323	420	301	289	347	344	579	363
theory	74	83	186	146	16	165	64	514	290
process	215	412	496	327	167	489	488	266	181

Table 4. Positive Specificities (nouns) taken from the English corpus

The presence of polysemic words in the list retrieved by the specificity is not surprising since the test relies solely on frequency and such words stand a better chance of having higher frequencies. Our corpora not being semantically tagged, we

are not able to automatically address the ambiguity. The manual description of the TLS will handle such cases and we will be able to distinguish meanings and discard non-relevant ones. The table also contains words like *particle* and *slave*, which are clearly domain related even though they are highly specific to our corpus. These will be handled by the next filtering step if needed.

Since we use the polarity of the specificity test, it is important to look at both ends of the spectrum to see what such a filter is discarding. Table 5 lists items that have been isolated by our statistical measure as significantly absent from our English scientific corpus. Once again a look at the table shows that the specificity test is working as expected and discarding data that should not be included in a TSL since most of the lexical items listed are either domain specific or closer to items of Basic English (Ogden 1930).

Word	Arch.	Chem.	Law	Geo.	Hist.	Comp. Sci.	Eng.	Phy.	Psy.
number	365	521	309	310	247	716	350	417	414
year	111	16	133	111	210	6	64	8	64
id	3	1	6	0	1	35	1	2	0
time	570	329	497	424	491	672	822	737	475
cent	8	2	10	91	33	0	7	1	5
subject	39	22	65	40	41	112	83	29	25
medium	16	43	9	9	7	12	57	17	27
government	59	0	826	239	633	16	1	0	2
yesterday	0	0	0	0	2	0	0	0	0
way	240	92	272	320	217	221	126	253	201
week	36	0	5	6	30	7	20	1	28
home	132	1	15	114	89	4	0	1	69
day	102	3	27	65	139	14	10	9	60
game	17	0	10	19	11	116	0	2	72
team	54	3	2	5	2	2	9	1	62
world	85	22	98	183	200	74	27	24	57
company	337	5	135	153	41	2	26	1	13
man	37	2	31	52	109	0	1	0	34
illustration	5	9	10	2	7	4	20	6	8
night	24	1	2	17	54	0	4	2	109
life	137	6	66	194	188	4	23	12	117
today	67	10	40	93	40	15	31	18	19
season	53	0	0	37	8	0	10	1	8
money	86	0	29	30	106	0	9	1	13
city	115	0	27	718	289	1	1	0	3
business	73	1	194	425	115	5	11	0	7

Table 5. Negative Specificities (nouns) taken from the English corpus

We must then look at the specificity threshold to determine what we consider to be a positive specificity. The last value can be linked to a probability of observing the frequency in the SC computed taking into account the frequency of the same word in the RC. Tables 6 and 7 show that applying a specificity based filter has an important impact on the number of lexical items to be considered as potentially part of the LST. As expected, when using a smaller probability threshold (p=1/1000), the number of items retrieved is significantly lower than with a larger threshold (p=1/100). The

reduction is similar in both languages with 34.5% less items retrieved with the lower probability for English and 37% for French.

Part of speech	Total	p=1/100	p=1/1000
Adjectives	49866	2642	1706
Adverbs	3675	587	400
Nouns	91020	4259	2886
Verbs	4735	711	430

Table 6. Specificity filtering on the English corpus

Part of speech	Total	p=1/100	p=1/1000
Adjectives	21883	2487	1594
Adverbs	1381	362	268
Nouns	74719	5682	3579
Verbs	41812	3905	1984

Table 7. Specificity filtering on the French corpus

This first step of filtering allows us to identify words that are very specific to the content of the corpus described, thus isolating our first subset of candidates for the description. One large discrepancy between the results in French and English is worth mentioning. It is interesting to note that the distribution of data is very similar in both languages for all parts of speech with the exception of verbs. We believe that the explanation for this phenomenon lies with the different tagging schemes used by the TreeTagger for English and French and our processing of the subsets produced by the tagger. The tagset used for French uses a finer grain description of the verbs and adds information about the tense to the lemma, which is not represented in the same way in the English tagset.

3.2. Distribution

In this section, we will describe how we can further filter the results using the distribution criterion. As can be seen in Tables 4 and 5, although some specific words are fully distributed throughout the corpus (*model*, *analysis* and *function*, etc.), some are partially distributed (*density*, *fig*, *particle*, *slave*, *and temperature*) even though they still appear in quite a few of them. The idea behind our distribution filtering is to keep solely lexical items that appear in all sub-corpora, regardless of areas. While the specificity test acknowledges the "scientific" aspect of the TSL, the distribution filtering step covers the "transdisciplinary" concept.

Tables 8 and 9 contain the most specific lexical items that are fully distributed in the French and English corpora. Most of the items that appear in the table seem relevant to our objectives. As can be expected from the specificity filtering, frequencies in each sub-corpus are mostly high with just a few exceptions. Examples of low frequencies are *siècle* (History) and *paramètre* (Law and Geography) for French or *phase* (History) and *surface* (Law and Hisory) for English. For the moment, these items are considered as part of the LST as they meet both filtering criteria of specificity and

distribution. However, one might wonder whether a third criterion of minimal frequency in each subcorpus should be added to our process.

Word	Arch.	Chem.	Law	Geo.	Hist.	Comp. Sci.	Eng.	Phy.	Psy.
type	2522	620	246	183	166	707	493	409	357
modèle	658	354	183	282	255	1785	1077	514	551
fonction	502	371	317	262	214	1559	651	893	725
phase	400	730	49	53	43	270	700	672	204
objet	1182	40	377	159	146	1124	110	159	1705
paramètre	56	367	5	11	10	207	614	545	20
contexte	528	16	102	159	83	1907	53	32	198
donnée	788	173	83	118	141	919	323	178	160
valeur	478	732	236	255	249	710	962	1133	399
élément	1048	143	409	131	172	714	318	319	302
structure	752	571	236	301	111	317	426	317	201
cas	1310	717	865	288	540	974	844	1036	456
profil	1078	106	2	22	45	426	133	148	61
effet	1450	690	989	293	377	440	552	954	854
siècle	2578	16	152	255	1346	10	4	21	118
section	678	19	131	61	10	510	303	250	30

Table 8. French fully distributed nouns with positive specificity

Word	Arch.	Chem.	Law	Geo.	Hist.	Comp. Sci.	Eng.	Phy.	Psy.
model	66	567	230	244	39	786	758	711	623
analysis	339	337	155	392	52	373	419	246	384
function	66	252	73	70	33	324	420	346	489
phase	83	419	40	31	8	127	227	785	291
system	183	489	1454	316	266	742	1132	1121	247
structure	126	469	92	183	59	277	411	341	117
method	59	303	58	80	49	380	377	175	210
state	122	308	912	360	417	379	119	1036	62
design	25	127	109	71	20	364	1016	288	156
interaction	36	173	17	148	13	126	218	253	213
research	378	107	769	498	21	327	283	46	476
surface	205	656	8	131	9	29	462	279	22
order	169	323	420	301	289	347	344	579	363
theory	74	83	186	146	16	165	64	514	290
process	215	412	496	327	167	489	488	266	181

Table 9. English fully distributed nouns with positive specificity

Applying a second filter to the results obtained from the previous specificity filter (with p=0,001) has a drastic impact on the number of items to be included in the TSL. Overall, 58% of the items are retained in English and about 48% for French when excluding verbs from the count (see 3.1 for a discussion).

Part of speech	p=0,001	Distrib=9
Adjectives	1706	381
Adverbs	400	170
Nouns	2886	551
Verbs	430	172

Table 10. Distribution filtering on the English corpus

Part of speech	p=0,001	Distrib=9
Adjectives	1594	338
Adverbs	268	135
Nouns	3579	612
Verbs	1984	684

Table 11. Distribution filtering on the French corpus

3.3. Overall results

What cannot be seen from the previous tables is a comparison of the final results when applying the distribution filter and both probability levels and the specificity filter.

Parts of speech	English	French
Adjective	397	354
Adverbs	177	146
Nouns	566	632
Verbs	184	750
Total	1324	1882

Table 12. Part-Of-Speech distribution of specific forms (p=0,01)

Parts of speech	English	French
Adjective	381	338
Adverbs	170	135
Nouns	551	611
Verbs	172	684
Total	1274	1768

Table 13. Part-Of-Speech distribution of specific forms (p=0,001)

If we compare the results of using a threshold of p=0,01 (Table 12) compared to one of p=0,001 (Table 13) for both languages, the amount of data to be validated by our team with the lower probability is not that much larger. We believe that we should start our description of the TSL using the larger set of results and manually exclude potential errors. Lexicographical descriptions are now under way for both languages and polysemic forms are being divided into different entries and described in an XML structure.

4. Future Work

The first step to be taken in the near future is to look at the data being discarded by both filtering steps, especially the borderline data having a p=0,01 and a distribution of 8 to see if any items that should be included in the TSL is being missed. We also need to investigate a third filtering step that would involve setting a minimal frequency threshold in all sub-corpora. As explained in section 3.1, the difference in the tagset between English and French had an important impact on the extraction of verbs. Verb tenses for French should be merged or aligned to the English tagset in order to see what this discrepancy implies.

We are now carefully looking at the results in order to manually distinguish word meanings in each language and to add definitions to the TSL. All data is contained in XML structures and will be published to the Web in the near future. A mapping mechanism from our format to the Lexical Markup Framework (LMF) format (Francopoulo *et al.* 2006) would be a good way to make sure the data can be reused in natural language processing research and distributed easily.

Once the language specific analysis and lexicographical descriptions are completed, we will be able to link the two lists together in order to present a bilingual TLS. Tables 8 and 9 already show some similarities between languages for the top most specific and distributed words. Data is now being handled separately in each language to avoid inter-language interference or influence. Moving from a monolingual description to a bilingual description will allow us to compare the results obtained in both languages. We expect to see discrepancies between languages. An analysis of these discrepancies will tell us whether we need to complete the data in one language or in another because some lexical items were left out or if the scientific language in English and French simply behave differently.

5. Conclusion

Based on the analysis and the partial validation of the data done to this date, we can state that the specificity and distribution are simple but effective clues for the extraction of a transdisciplinary scientific lexicon. Using more sophisticated algorithms and natural language processing tools might lead to faster and more precise results but the amount of data retrieved is small and is, in most cases, relevant. We believe that dividing forms extracted into various lexical units based on their meaning is still work that should be taken care of by a lexicographer. Using semantic tagging right from the start and fixing the output of such tools might lead to a lengthier processing than our approach.

Although it could be applied to various areas, our work is performed within the study of language for special purposes. It is our opinion that more applied studies need to be done on this type of language, especially in relation with terminology descriptions. The exploration and the description of “lexical layers” or “lexical subsets” that can be found in scientific documents would be beneficial to areas like language teaching and learning, terminology description and to natural language processing.

References

COXHEAD, A. (1998). *An academic word list*. Wellington: Victoria University of Wellington.

COXHEAD, A. (2000). A New Academic Word List. *TESOL Quarterly*, 34(2): 213-238.

FRANCOPOULO, G., GEORGE, M., CALZOLARI, N., MONACHINI, M., BEL, N., PET M. and SORIA C. (2006). Lexical markup framework (LMF). In *Proceedings of LREC*: 223-236.

LAFON, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus, *MOTS*, 1: 128-165.

LEE, D. (2001). Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, 5(3): 37-72.

OGDEN, C.K. (1930). *Basic English: a general introduction with rules and grammar*. London: Kegan Paul, Trench, Trubner.

SCHMID, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

TUTIN, A. (2007). Présentation du numéro Autour du lexique et de la phraséologie des écrits scientifiques. *Revue française de linguistique appliquée*, 12(2): 5-13.

WEIR, G. and N. ANAGNOSTOU (2007). Exploring Newspapers. A Case Study in Corpus Analysis. In *Proceedings of ICTATLL 2007*.

WEST, M. (1953). *A General Service List of English Words*. London: Longman.

Specialised lexicographical resources: a survey of translators' needs[1]

Isabel Durán-Muñoz[2]
University of Málaga

Abstract

Lexicography is facing new challenges in the 21st century and therefore, new and more appropriate applications are being developed to satisfy users' needs and adapt to new technologies. But how satisfied are users with lexicographical resources? Which level of satisfaction is reached by this kind of resource? These questions have already been answered by different scholars regarding different types of users, but what happens with translators? In our opinion, professional translators have always been pushed into the background and therefore, there is a lack of concrete and useful information about them as real users.

In this paper, we present a survey carried out to improve the development of lexicography regarding professional translators' needs and expectations about specialised lexicographical resources. This project seeks to fill this existing gap by identifying the real needs of translators with regard to terminology. More specifically, we present the results of a recent survey in which translators were asked which terminological resources they currently use and what resources they would ideally like to use, in order to identify their expectations and desiderata about their "ideal" terminological resource prior to the development of such a resource. It is hoped that the identification of users' needs with regard to terminology could lead to useful resource development projects in the future.

Keywords: specialised lexicographical resources, professional translators, survey.

1. Introduction

It is an unfortunate reality that the majority of resources currently available are of little use to translators, and therefore many are obliged to resort to the creation of their own terminological resources either from comparable corpora or from existing translations. These inadequate resources often pose a problem for translators since it is well known that they usually work under time pressure and they do not have the opportunity to create their own resources. This is the reason why terminological resources have considerable importance for them and should meet their requirements as far as possible. Unfortunately, these resources are frequently of poor quality and do not adequately satisfy their needs.

In our opinion, this reality is due to the fact that professional translators have always been pushed into the background and their needs for this kind of resource have hardly

[1] The research reported in this paper has been carried out in the framework of project BBF2003-04616 (Spanish Ministry of Science and Technology/EU ERDF).

[2] iduran@uma.es

ever been seriously taken into consideration. This is the reason why most terminological resources do not fulfill their expectations, both regarding quality and quantity. Therefore, professional translators are frequently obliged to create their own terminological resources, either from *ad hoc* corpora or translation memories.

Moreover, surveys or research conducted so far in relation to lexicographical and terminological resources seem to have been limited to foreign language or translation students and their ability to look up definitions in dictionaries (Bejoint 1981, Roberts 1992, Duvå and Laursen 1994, Dancette and Réthoré 1997, Mackintosh 1998, Varantola 1998, Hartmann 1999, Corpas Pastor *et al.* 2001, Sánchez Ramos 2005, Bogaards 2005, East 2008) but none of them focused on professional translators.[3] In this sense, there is a lack of concrete and useful information about these users, who present a number of specific features and needs regarding these resources.

This study set out to investigate how professional translators use terminological resources and which necessities and difficulties they encounter when they use them. The intention was to provide some insight into professional translators' look-up processes and to examine their needs and expectations, as well as to identify the existing gap between translators' real needs and expectations and the information contained in this kind of resource.

2. Previous studies about resources and users' needs

A number of studies and research about terminological resources and users' needs has been carried out in the last decades, aiming to find out either the adequate content for their potential users or the skills needed to properly use these resources. However, in these studies, to the best of our knowledge, there was no interest in focusing on professional translators as real users and thus, these were ignored and not taken into consideration, although their work is mainly based on the use of this kind of resource.[4]

The majority of these previous studies have been based on foreign language students or translation students but none of them on professional translators. In our opinion, professional translators must be considered as a concrete and different group of users, since they need specific terminological resources to carry out their work and thus, they require concrete information to satisfy their look-up needs. In this sense, they should be considered as real users and therefore, be taken into account during the preparation phase of a terminological resource and offered specific resources.

Previous studies that focus on dictionary use can be classified in two main groups: on the one hand, those which study appropriate skills in dictionary use; and on the other hand, those whose aim is to identify users' (specifically translators') needs and expectations on dictionaries.

[3] In their study, Duvå and Laursen (1994) worked with a group of informants who were partly graduates and professional translators (the latter accounting for 38% of the total).

[4] The documentary phase (above all consultation of terminological resources) in the translation process occupies more than half of the time for the translator.

The first group – studies about dictionary use skills – claims it is necessary to teach certain skills to students in order to improve the use of lexicographical resources. These studies consider a dictionary as a special book that, in order to be effectively used, requires certain abilities to find the information being sought, *i.e.* users need specific training for the use of these resources. The authors claim that, used appropriately, the dictionary can be an invaluable tool for learners of a foreign language; but without proper skills the dictionary can be as much of a hindrance as an aid. It seems, however, that many users lack appropriate skills and hardly receive any dictionary training.

Works related to studying how users perceive and use dictionaries have been mainly focused on learners of second languages (Bejoint 1981, Hartmann 1999, Bogaards 2005, East 2008), although we can also find some research about trainee translators (Roberts 1992, Atkins and Varantola 1998, Mackintosh 1998, Varantola 1998, Sánchez-Ramos 2005).

Regarding studies focused on translation students, a number of scholars (Roberts 1992) claims that translators as language users need to know how to effectively consult and use dictionaries in order to complete the translation process with success. So, to them, it is essential to further study the relationship between trainee translators and specialised dictionaries. As a result, they carry out empirical research on habits of use, needs and different problems that dictionaries can cause to students.

The second group mentioned above – studies about translators' needs and expectations – is closer to our research. As we said above, the previous studies carried out regarding this topic were all focused on translation students (Duvå and Laursen 1995, Dancette and Réthoré 1997, Corpas *et al*. 2001). Up to now, we have not found any study about professional translators' needs and expectations. In order to solve this gap, we carried out our study, which will be described in the following sections.

The aim of these previous studies was to identify the resources that translation students use when they are translating a text and the necessities or difficulties that they encounter during the process, *i.e.* which information they consult (grammar, definition, etc.) and where they look it up, which difficulties they encounter when they are consulting a specific term or construction, among others.

Our goal is similar but the recipients are different. We intend to find out what and how professional translators consult terminological resources and which problems or difficulties they encounter when they do so. We are also interested in identifying the type of information they would like to find in a resource of this kind.

All the previous studies concluded that translators – without distinguishing between students and professionals – require the following information: linguistic information (*e.g.* definitions), semantic information (*e.g.* semantic relations), and pragmatic information (*e.g.* context).

In our study, we will try to find out if these requirements are also demanded by professional translators or if, on the contrary, they need some different data.

3. The survey: description and results

Terminological works always start from a study about the potential users of a resource project, in order to know which needs they have, what they expect, and which information they do (not) require. According to Stein (1984: 4):

> Dictionaries are obviously written for their users. We therefore need much more research on the dictionary user, his needs, his expectations, and his prejudices.

Bergenholtz and Tarp (1995: 77) also point out the necessity of carrying out previous communication with the potential user before starting the terminological work so as to include, or exclude, specific information.

> Lexicographical work often proceeds without any prior knowledge of the potential user group, and the dictionary may therefore be said to be the result of the lexicographer's own conjectures concerning user needs for lemmata, collocations, sentence examples, encyclopaedic and linguistic information, etc. To acquire more precise knowledge, the lexicographer may make a user survey before starting actual work on the dictionary, with the aim of uncovering the needs of potential users in relation to the information categories to be incorporated in the dictionary as well as the representation of this information.

For professional translators, this research is absolutely essential if we take into account that translators spend a substantial amount of time and effort consulting these sources (Varantola 1998). In other words, professional translators need to be considered as real users and then, be asked which needs and expectations they have.

These studies about users can be carried out by employing different empirical techniques: protocol techniques, which are characterized by the fact that informants, at the same time as they are carrying out a particular activity, register exactly what they are doing (Atkins and Varantola 1998, Duvå, G. and Laursen 1995); personal interviews, which may be applied either on their own or in combination with other methods and consist of personal interviews with informants, who are asked direct questions about a previous task or about their experience (Duvå and Laursen 1995); or questionnaires/surveys, by which informants answer some previously defined questions about a specific topic, which can be very diverse (multiple choice, yes/no questions, etc.) (Corpas *et al*. 2001).

These three methods present both advantages and disadvantages, and they must be selected according to the requirements of the research. We preferred to employ the survey method due to the advantages it presents against the other techniques.

The main advantage of this technique is the possibility of reaching a very large population in a very short time, which is not possible with other empirical methods. Also, it can be administered from remote locations using mail, email or telephone; it is feasible to make more precise measurement by enforcing uniform and comparable answers and automatic quantitative analysis. Another basic advantage is that filling in survey questions is less time-consuming than other empirical methods, which is very important taking into account that our recipients (professional translators) do not have a lot of time to waste filling in surveys.

Nevertheless, it is also important to take the disadvantages into account and try to minimise their negative effects: lack of interest and participation, and lack of reliability in answers. These two problems can be reduced if the target population is well established and the channels of distribution are also well selected. In our case, these two disadvantages were to some extent eliminated by focusing on professional translators as the target recipients and by selecting the professional associations, organisations, etc. as the distribution channels.

3.1. Preparation and description of the survey

This survey was designed in line with recent established survey practices (Dillman 2007, Groves *et al*. 2008) and launched in July 2008 in English, Spanish, Italian and German. It was activated during the following three months[5] and sent to professional translators via specialised mailing lists (Corpora List, The Linguist, Termilat, Traducción. among others), and through several organisations for translators and interpreters (ACT, AIETI, ASETRAD, ITI, ASTTI, etc.). It was also sent out to a number of translation companies as well as individual translators. These contacts were not limited to one country, or several countries, but to organisations, companies, translators, etc. around the world.

The survey was addressed to all types of translation professionals (translators, terminologists, project managers, subtitlers, etc.) (see Figure 1 below). Its main goal was to shed light on their opinion about the current terminological resources and on their use of these resources and their needs while translating. Moreover, we aimed to obtain information about the different terminological resources they used and their preferences regarding content and organisation.

In total, 402 answers were obtained during the period the survey was open, from which can be drawn conclusions on the elaboration of terminological resources for translators in any specialised domain.

During the preparation phase, several important issues had to be carefully considered. For instance: how can one get information on what the users need and expect? Or how can the researcher be sure about what participants understand from the question? In order to minimise any misunderstanding, ambiguity or loss of information due to the issues above, a pilot study was carried out prior to the survey being completely designed. This previous study was addressed to domain-related experts, *i.e.*, experts in translation and terminology, aiming to enhance the initial version of the survey and to elaborate a survey which covered the proposed necessities and goals of this study. To do so, these experts were contacted through e-mail and asked to fill in the survey and give some feedback (comments, recommendations, proposals, etc.). Once the feedback was received, the appropriate changes were made and the final version was completed. Hence, a high quality survey was obtained where all the relevant questions were included in a clear, simple and direct way.

[5] The survey link was http://clg.wlv.ac.uk/surveys/survey.php?sid=29

The survey consists of 20 questions in total, classified in 4 different sections: 1. Professional information; 2. Working environment; 3. Terminological resources, and 4. Assessment of resources used by translators and their views on 'ideal' resources.

The first part of the survey (first two sections) was focused on the characterisation of participants, so as to obtain information about different aspects of their academic and professional experience. The informants were asked to provide information about their background (education, profession, experience) and about their working environment (working languages, domains/genres that they usually translate, use of internet). The second part focused more specifically on various aspects of terminological resources: users were asked to identify the terminological resources that they use (encyclopedias, dictionaries, thesauri, parallel corpora/texts, etc.), the format of these resources, the organisation structure they prefer, etc. Finally, users were asked about their own assessment of the resources. In doing so they were asked to consider any problems or inconveniences they may have experienced, taking into account issues of presentation and information they thought should be present in an “ideal” terminological resource.

3.2. Results of the survey

In order to briefly illustrate the participants’ profile, we will present some general information obtained in the first part of the survey.

The respondents who declared themselves to be translators totalled 62.55%, 13.67% were interpreters, 6.67% project managers, 5.24% terminologists, 3.00% subtitlers, and 8.80% worked in another profession (see Figure 1).

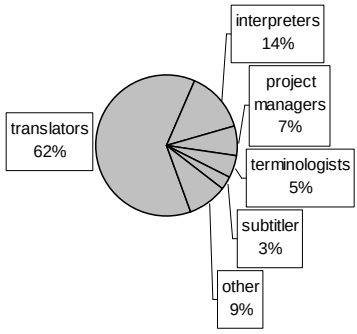


Figure 1. Participants

The majority indicated that they held a degree in Translation and Interpreting (61.96%), or at least had some studies in this domain (Translation Memory courses, specialised translations, etc.), and only 5.08% reported not having a professional qualification relevant to their job. Their professional experience was more than 10 years in 41.65% of cases, and their working language was mainly English in any type of specialised domain (32.12%), followed by Spanish (15.80%), French (15.21%),

German (9.94%) and Italian (6.80%).[6] Regarding the most common working specialised domain,[7] the volume of translations in the legal domain occupies the first position with a 36.57%, followed by business translations (34.82%). Next, we find translation in the ICT domain (30.35%), humanities (28.11%), and arts, literature or media with 27.6% (see Figure 2).

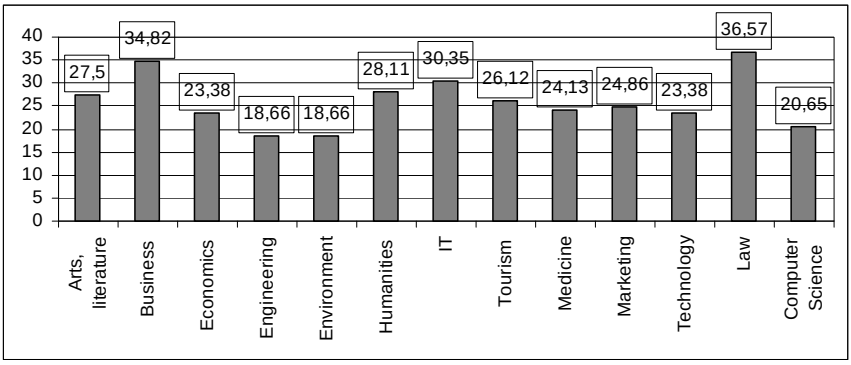


Figure 2. Specialised domains with a percentage over 20%

The question regarding the specialised-working domains was not a restricted one and translators were able to select several specialised domains, which gave us an idea about the number of different domains in which they work. According to these results, the average number of different working domains for translators is six, although there are also professionals who only work in one domain or who work in more than ten. At the same time, some professionals work in close domains, such as business, mathematics/statistics, economics and marketing, but others work in very different domains, *e.g.* law, handwork, geography and ICT.

Having presented the participants' features, we now describe their preferences and their needs regarding terminological resources.

The participants prefer online resources (56.47%) to any other type (electronic resources accounted for 24.71% and paper-based 18.82%) mainly due to easy and quick access. This fact indicates that professional translators need high-quality online terminological resources so as to provide good results in their translations. However, most of the online resources currently available obtain a very poor qualification against electronic or paper-based resources in a study carried out to assess this kind of tool.[8]

[6] Apart from these four languages, translators were able to select 36 other languages, among them Greek, Portuguese, Russian, Chinese, Ukrainian, etc.

[7] The specialized domains included in the survey were taken from the British Institute of Translation and Interpreting (ITI. URL: http://www.iti.org.uk/indexMain.html).

[8] The study for the assessment of these resources has also been carried out within the framework of the project BBF2003-04616 (Spanish Ministry of Science and Technology/EU ERDF), but it has not been published yet.

Translators prefer bilingual (39.45%) to monolingual resources in the target language (25.56%) and source language (24.12%), and above all to multilingual resources (10.88%). This information is relevant to terminographers, since the most convenient resources for these users prove to be bilingual resources, or at least monolingual, rather than multilingual. This is due to the fact that users consider multilingual resources as having lower quality and being less useful in their work.

To the question, "Which type of terminological resources do you use most when translating?," the participants selected from the list[9] included in Table 1.

Bilingual specialised dictionary/glossary	18.94%
Searches in search engines (*Google*)	16.13%
Terminological databases	8.84%
Monolingual specialised dictionary/glossary (L1)	8.63%
Wikipedia	8.63%

Table 1. "Which type of terminological resources do you use more when translating?"

In this table, only the first five resources are shown. According to these results, the preferred resources are specialised bilingual dictionaries (18.94%), followed by searches in search engines like *Google* (16.13%). The third position is occupied by terminological databases (8.84%), followed by monolingual dictionaries (L1) (8.63%), and Wikipedia (8.63%). In the following positions, we find other resources, such as monolingual specialised dictionaries (7.83%) or parallel corpora (5.09%), but with a lower percentage.

Here we observe some unexpected data regarding Wikipedia, since, according to this information, this resource occupies the fifth position among the preferred resources by professional translators. Hence, it must be considered as a common and frequently used resource by these users[10] and therefore, as a new possibility of searching information, despite the negative criticisms that this resource has received due to its doubtful reliability.

With regard to the basic criteria to assess currently available terminological resources, 38.71% of the participants indicated that they do not use any resource if this is not reliable, against 33.71%, who considered that it is not always possible to find reliable resource and 19.11%, who acknowledged that they did not carry out any previous assessment of the resource.

[9] The complete list of resources and their percentages is given in Appendix 1.

[10] The percentage obtained by Wikipedia is low (8.63%), but here we must take into account that participants had to choose among a wide range of resources and this is why none of the percentages are very high. In Appendix 1, the total selections for each resource are shown in the first column.

In order to identify the most relevant criteria employed by the participants to determine the reliability and quality of resources, they were asked to order seven different assessment criteria, from 1 to 7, where 1 was the most relevant criterion of selection. The criteria ordered according to their preferences were the following: 1. Authorship; 2. Specialisation of the website; 3. Richness of information; 4. How up to date it is; 5. Ease of access; 6. External comments about the resource, and 7. Instructions for use. According to these results, professional translators consider authorship as the most important criterion of reliability, followed by the specialisation of the website, and as the least relevant the external comments about the resource and the inclusion of instructions for use.

We will now present the results of the key question of the survey: "What do you think a good terminological resource for translators should offer?" The different options were based on categories included in ISO 12620:1999. The results differed somehow but in general they coincide in the information which should be considered essential, desirable and irrelevant. With these results, terminographers of resources for translators will have a very clear idea about which information translators need, prefer and expect and what is irrelevant to them. Also, terminographers can identify the information that is not considered as essential by translators but which can be interesting to include in a resource targeted at them, *i.e.* desirable data.

Essential data	**Desirable data**	**Irrelevant data**
Clear and concrete definitions	A great variety of units (n., v., adv., adj.)	Etymological information
Equivalents	An explanation of each translation equivalent	Pronunciation
Derivatives and compounds	A greater variety of examples	Syllabification
Domain specification	Grammatical information	
Examples	Semantic information (semantic relations, frames)	
Phraseological information	Pictorial illustrations	
A definition in both languages (if bilingual) (45.11%)	A definition in both languages (if bilingual) (45.38%)	
Abbreviations and acronyms	Instructions for use	

Table 2. What do you think a good terminological resource for translators should offer?

From table 2 we can draw some conclusions about what professional translators need and expect from a terminological resource. Undoubtedly, these users require information that helps them to codify the new message, that is: on the one hand, linguistic information (definitions, equivalents, collocations, acronyms, etc.); and on

the other hand, pragmatic information (domain specification, context). The rest is desirable but not essential, *i.e.* semantic information, images, grammatical information, etc.

There is only one option that is repeated both as essential and desirable data, which is "definition in both languages". The fact that the percentages are so similar (45.11% and 45.38%, respectively) makes it clear that professional translators are not in agreement on this point. Therefore, it is up to the terminographers to decide to provide both or just one definition in their terminological resource for translators.

To conclude, the survey also offered the opportunity to give some feedback through an open question ("Do you have any other suggestion about the content of a good terminological resource for translators?"). The answers given were very interesting and the following ones recurred most:

- Exportability (.txt or .tmx)
- Clarifications and examples about use (the translations that should NOT be used because they are tricky, inconvenient, false cognates, etc.)
- Information on example sources (references, URLs, etc.)
- Cultural differences between source and equivalent term, and regional variations
- Links to other resources.

These are options that are not usually included in this kind of resource but are required by these specific users. Consequently, it is clear that professional translators have hardly ever been taken into account as potential users in the design of terminological resources.

4. Conclusions

The results obtained in this research clarify the needs and expectations that professional translators as real users have. We now know more about their opinions regarding the current terminological resources and have given them the opportunity to describe their "ideal" resource.

Translators are not real experts in the numerous and different domains they work in and thus, their translation process is mainly based on all the terminological resources they consult. Hence, they need appropriate resources including adequate information in order to satisfy their needs and thus, to be able to provide high quality results in their translations.

The results obtained from this research differ from the conclusions drawn in previous studies (based on trainee translators and second language students), which defended the needs to include linguistic, pragmatic and semantic information in resources for translators. Here we see that professional translators consider semantic information (semantic relations, semantic frames or domains, etc.) as desirable data but not as essential data, *i.e.* they do not see this information as essential but only as

complementary. Consequently, we observe how trainee translators need different information than professional translators and therefore, it is necessary to know their specific needs in order to elaborate resources for these professionals.

Also, it is relevant to take into consideration the needs of professional translators in order to provide quick and easy access to information in online resources and include good and concrete definitions together with pragmatic information (context, tips of use, information about false friends, etc.) which will help them understand the source term and correctly translate it.

References

ATKINS, B.S. and VARANTOLA, K. (1998). Monitoring dictionary use. In B.S. Atkins (ed.). *Using Dictionaries*. Tübingen: Niemeyer: 83-122.

BEJOINT, H. (1981). The foreign student's use of monolingual English dictionaries: A study language needs and reference skills. *Applied Linguistics*, 2(3): 207-221

BERGENHOLTZ, H. and TARP, S. (1995). *Manual of specialised lexicography: the preparation of specialised dictionaries*. Amsterdam, Philadelphia: John Benjamins.

BOGAARDS, P. (2005). A propos de l'usage du dictionnaire de lange étrangère. *Cahiers de Lexicologie*, 52(1): 131-152.

CORPAS PASTOR, G., LEIVA ROJO, J. and VARELA SALINAS, M.J. (2001). El papel del diccionario en la formación de traductores e intérpretes: análisis de necesidades y encuestas de uso. In M.C. Ayala Castro (ed.). *Diccionarios y enseñanza*. Alcalá: Universidad de Alcalá: 239-273.

DANCETTE, J. and C. RÉTHORÉ. (1997). Le dictionnaire bilingue (anglais-français) de la distribution: entre dictionnaire de langue et encyclopédie. *Meta* XLII(2): 229–243.

DILLMAN, D. (2007). *Mail and Internet Surveys: The Tailored Design Method*. New York: John Wiley & Sons.

DUVÅ, G. and LAURSEN, A.L. (1995). Translation and LSP Lexicography: A User Survey. In H. Bergenholtz and S. Tarp (eds). *Manual of Specialised Lexicography. The preparation of specialised dictionaries*. Amsterdam, Philadelphia: John Benjamins Publising Company: 247-267.

EAST, M. (2008). *Dictionary Use in Foreign Language Writing Exams. Impact and implications*. Amsterdam, Philadelphia: John Benjamins.

GROVES, R., FOWLER, F., COUPER, M., LEPKOWSKI, J., SINGER, E. and TOURANGEAU, R. (2004). *Survey Methodology*. New Jersey: John Wiley & Sons.

HARTMANN, R.R.K. (1999). Case study: the Exeter University survey of dictionary use [Thematic Report 2]. In R. Hartmann (ed.). *Dictionaries in Language Learning*. Berlin: Thematic Network Project in the Area of Languages: 36-52.

MACKINTOSH, K. (1998). An empirical study of dictionary use in L2-L1 translation. In B.S. Atkins (ed.). *Using Dictionaries*. Tübingen: Niemeyer: 123-149.

ROBERTS, R.P. (1992). Translation pedagogy: strategies for improving dictionary use. *Traduction, Terminologie et Rédaction*, 5(1): 49-76

SÁNCHEZ RAMOS, M.M. (2005). Research on Dictionary Use by Trainee Translators. *Translator Journal*, 9(2). URL: http://accurapid.com/journal/32dictuse.htm [Consulted on 15/12/2009].

STEIN, G. (1984). *The English dictionary: past, present and future*. Special lecture given at the inauguration of the Dictionary Research Centre. Exeter: Exeter University Press.

VARANTOLA, K. (1998). Translators and their use of dictionaries. In B.S. Atkins (ed.). *Using Dictionaries*. Tübingen: Niemeyer: 179-192.

Appendix 1:

Monolingual specialised dictionary/glossary (L1)	129	8.63%
Monolingual specialised dictionary/glossary (L2)	117	7.83%
Bilingual specialised dictionary/glossary	283	18.94%
Multilingual specialised dictionary/glossary	42	2.81%
Monolingual visual dictionary	8	0.54%
Bilingual visual dictionary	20	1.34%
Section Images in a search engine (like Google)	53	3.55%
Searches in search engines (like Google)	241	16.13%
Parallel corpora (original texts and their translations)	76	5.09%
Comparable corpora (original texts in both languages)	69	4.62%
Terminological database	132	8.84%
Encyclopaedia	46	3.08%
Wikipedia	129	8.63%
Mailing lists	24	1.61%
Internet forum	62	4.15%
Thesaurus	48	3.21%
Other	15	1.00%

DIL: a German-Italian online specialized dictionary of linguistics

Carolina Flinz[1]
University of Pisa

Abstract

DIL is a bilingual (German-Italian) online dictionary of linguistics. It is still under construction and contains 240 lemmas belonging to the subfield of "German as a Foreign Language", but other subfields are in preparation. DIL is an open dictionary; participation of experts from various subfields is welcome. The dictionary is intended for a user group with different levels of knowledge, therefore it is a multifunctional dictionary. An analysis of existing dictionaries, either in their online or written form, was essential in order to make important decisions for the macro- or microstructure of DIL; the results are discussed. Criteria for the selection of entries and an example of an entry conclude the article.

Keywords: online dictionary, open dictionary, linguistics, bilingual, German-Italian, German as foreign language.

1. Introduction

This paper describes some selected aspects of an online German-Italian specialized dictionary covering the field of linguistics. Following the University Reform of 1999, the need for such a tool in Italy was particularly strong, as no such dictionary existed either in a printed version or online.

DIL (*Deutsch-Italienisches Fachwörterbuch der Linguistik*) is a project of the Institute of Linguistics of the University of Pisa (Foschi Albert and Hepp 2004: 37). In order to make it available to as many users as possible it was created as an online dictionary with free access via the internet. It was published in 2008 on the University's server (http://www.humnet.unipi.it/dott_linggensac/glossword/), and is regularly updated.

DIL is a bilingual dictionary (German-Italian), but it is "monolemmatized" in the sense that it is one-way: the entries are in German, but the equivalents and the explanations are in Italian.

It is a specialized dictionary covering the field of linguistics. At the moment only the subfield of DaF (*Deutsch als Fremdsprache, i.e.* "German as a foreign language") is

[1] carolinaflinz@virgilio.it

complete, but other subfields (such as morphology, lexicography, etc.) are either in preparation or planned.

DIL is an open dictionary, so the participation from experts from the various subfields is welcome. The concept is that of "Wikipedia", but the entries are strictly controlled and revised by an academic committee before publication. The author remains responsible for the entries and can be contacted by the users.

After stating the specific functions of the dictionary, the needs of the potential user group and its possible usage (Section 2), an analysis of the existing dictionaries within this field, both in their written and their online forms (Section 3), was carried out. These analyses were of great importance in order to create a dictionary that would combine both the qualities of a print and of an online dictionary. This helped to provide guidelines for the composition. The lexicographic basis and the criteria for the compilation are also briefly presented (Section 4). An example of an entry taken from the dictionary (Section 5) concludes the article.

2. Dictionary Functions

Dictionary functions are strictly related to the intended user group (a), its needs (b) and the potential usage situation (c). All of these were crucial for the design and preparation of this dictionary.

(a) The identification of the potential user is one of the primary requirements in planning a dictionary (*cf.* Barz *et al.* 2005: 15), but this is a recent interest in lexicography (*cf.* Hartmann 1983) and a real turning point (*cf.* Zöfgen 1991: 2896). Usually the lexicographer has a special set of users in mind from the beginning.

Even in the case of LSP dictionaries, which represent a category in its own right as they are designed for small, homogenous groups with similar characteristics, the potential user group is normally heterogeneous.

Lexicographers therefore need general guidelines for their project and recent literature distinguishes between three main user groups: experts, semi-experts and laypeople (*cf.* Nielsen 1990: 131, Bergenholtz and Kaufmann 1997: 98-99).

The intended user group of DIL is a mixture of these three categories, with different levels of knowledge and language competence. It encompasses both the layperson and the expert (student or teacher from various fields, such as germanistics, linguistics or German as a foreign language, authors of books, lexicographers or academics in general), so flexibility is important.

Lexicographers usually have to make a profile of the intended user group. DIL distinguishes between:

- a primary user group: users with Italian as their mother tongue and German as a foreign language;

- a secondary user group: users with German as their mother tongue and Italian as a foreign language;

(b) The needs of the intended user group can be assessed in many ways. Lexicographers usually differentiate between: listing of dictionary user habits, users' experiences, analyses of the real, concrete needs of the user group in a particular usage situation and listing of the hypothetical problems of the user.[2] The results determine what type of information should be included in the dictionary and what can be omitted.

Different methods can also be used (questionnaires[3], observation, experiments, using protocols, interviews etc.) in order to analyze the different needs. DIL used written questionnaires (sent by e-mail) and interviews.

Every method has a higher or lower level of objectivity: while more objective methods are preferred (*cf.* Baunebjerg Hansen 1990: 8), it is also important to include the introspection of the lexicographer (*cf.* Barz *et al.* 2005: 83). Through accurate analyses, lexicographers can avoid many critical comments (*cf.* Zöfgen 1994: 51). In order to have a concrete contact with the users and to try to fulfil their needs in the most effective ways, DIL has chosen to add a questionnaire that the user can read, save and answer.

(c) In order to have a real idea of the potential user group and of its needs, lexicographers recommend defining the possible usage situation (*cf.* Wiegand 1977: 101; Kühn 1983). Usually, lexicographers choose one of two possible situations:

1. The user needs some specific information and looks for it in the dictionary;

2. The user uses the dictionary in order to solve a problem (reception, production or translation).

DIL will try to fulfil both needs.

Dictionary functions represent a compromise between the needs of the user group and the information found in a dictionary in order to meet these needs.[4] A dictionary can perform many functions and an LSP Dictionary even more so, but multifunctionality is always recommended (*cf.* Bergenholtz 1992: 49). The distinction between more and less important functions is fundamental and helps the lexicographer in the definition of the lexicographic basis, the selection of entries, the choice of information following

[2] *Cf.* Wiegand (1977); Kühn (1983); Hartmann (1983); Baunebjerg Hansen (1990); Zöfgen (1991); Storrer and Harriehausen (1998); Schaeder and Bergenholtz (1994); Zöfgen (1994); Barz *et al.* (2005).

[3] The pros and cons of making use of questionnaires are discussed in Barz *et al.* (2005: 85), Ripfel and Wiegand (1988: 493).

[4] Many authors have investigated functions (*cf.* Sĉerba 1982, Hausmann 1977, Kromann *et al.* 1984, Mugdan 1992), trying to build classifications and typologies; we quote only some of them: communication and knowledge functions (*cf.* Schlaefer 2002); direct or indirect functions (*cf.* Tarp 1994: 230f); active and passive functions (*cf.* Sĉerba 1982: 52ff); productive and receptive functions (*cf.* Welker 2003: 12); "enkodierende und dekodierende Funktionen" (*cf.* Wiegand 1998); text dependent and text independent functions (*cf.* Bergenholtz and Kaufmann 1997: 98-99); L1-L2 or L2-L1 translating functions (Hausmann 1977).

the entry (Kromann *et al.* 1984: 167), the determination of the "outside matter", and the construction of the layout and web design (*cf.* Barz *et al.* 2005: 21).

DIL is a multifunctional dictionary and performs a plurality of functions, in that it is both active and passive, but also usable for the production and reception of texts. In addition, it can be used for translating from L1 to L2 and from L2 to L1.

On the basis of the intended user group and of their factual and linguistic competences (both in L1 and L2), the needs that DIL has to fulfil and its functions can be summed up as follows:

- The Italian mother tongue user can use DIL in order to perform the following operations: understand an LSP German word; translate a German word; learn more about an LSP word (the user wishes not only to find information but also to improve his encyclopaedic knowledge).
- The German mother tongue user will use DIL when translating into Italian; and/or when he/she is producing LSP texts in Italian.

3. Analyses of the existing dictionaries of linguistics

Two analyses proved to be of great importance in order to make decisions about the macro- and micro-structure of the dictionary: 1) the analyses of written monolingual dictionaries of linguistics; and 2) the analyses of online monolingual and bilingual dictionaries covering the field of linguistics.

1) The first important monolingual dictionaries of linguistics were published in Germany at the beginning of the 70s and in Italy at the end of the 80s. Before this period, there were only translations from French and from English. Even if the two countries had a similar evolution, the lexicographic work was much more intensive in Germany. In addition, the number of published dictionaries was higher and nowadays there are many more dictionaries in Germany (Bußmann, Lewandoswki, Metzler, etc.) than in Italy (Beccaria, Cardona). After 1999, we could not identify any publications in either country, the one exception being the Italian translation from Bußmann (2007).

Even though monolingual dictionaries differ from each other, in the sense that they have different entries, the given information is not the same and the structure of the entries changes. The analyses of a small corpus of dictionaries revealed that a written dictionary is made of at least the following components, also called "outside matter": a) an introduction; b) a register of the entries; c) a list of abbreviations; d) a guide for the correct usage of the dictionary; e) a bibliography.

While both German and Italian dictionaries have these components, the major difference concerns the bibliography: German dictionaries place a precise bibliography at the end of each entry, so that the users who wish to know more about the theme can find further information; Italian dictionaries have only a large and detailed

bibliography at the end of the dictionary. This bibliography is not of great use for the layperson or semi-expert user, who will tend to get lost in detail.

2) From the start of the 80s, many dictionaries of linguistics (both monolingual and bilingual) have been published online.[5] If we enter "dictionary of linguistics" in a search engine like *Google*, *Altavista* or *Lycos* the results include not only dictionaries but also lexicons and glossaries. One search result (April 2009) includes 24 dictionaries/glossaries[6]: 17 are monolingual, 5 bilingual and 2 plurilingual.

Some general considerations can therefore be made:

a) The term "dictionary" is used often in a general way, rather than following lexicographic criteria. The limits between dictionary, archive, grammar, databank are not strict and this can be seen as a typical characteristic of the online product (*cf.* Abel 2006: 52);

b) There is an extensive use of the terms "glossary" and "lexicon", often interchangeably;

c) There are both dictionaries that have been explicitly created for the online medium and dictionaries that are adaptations of written ones. We also found dictionaries that are added as Word or pdf files;

d) 83% of the published dictionaries have been financed by universities or university institutes; 8% also belong to the academic world, but they are published privately by university professors or research groups;

e) Many works were produced as a result of university courses, seminars, lessons or written publications;

f) Some dictionaries are linked to large portals;

g) Most of them have never been modified after their publication; updates are not often seen (only 25% have been updated since 2007);

h) Some dictionaries are still present on general portals but the links are no longer functioning so that they cannot be opened;

i) The copyright for most of them is shown through the symbol ©.

The analyses of the "outside matter" have shown a different situation when compared with written dictionaries, since there is much more flexibility in online usage.

[5] Compared to the number of dictionaries of the general language, the number of published dictionaries of linguistics is very small.

[6] Only dictionaries of general linguistics have been listed. Specific dictionaries concerning only one subfield or similar works have been excluded. It was noted that there is a large number of grammatical dictionaries.

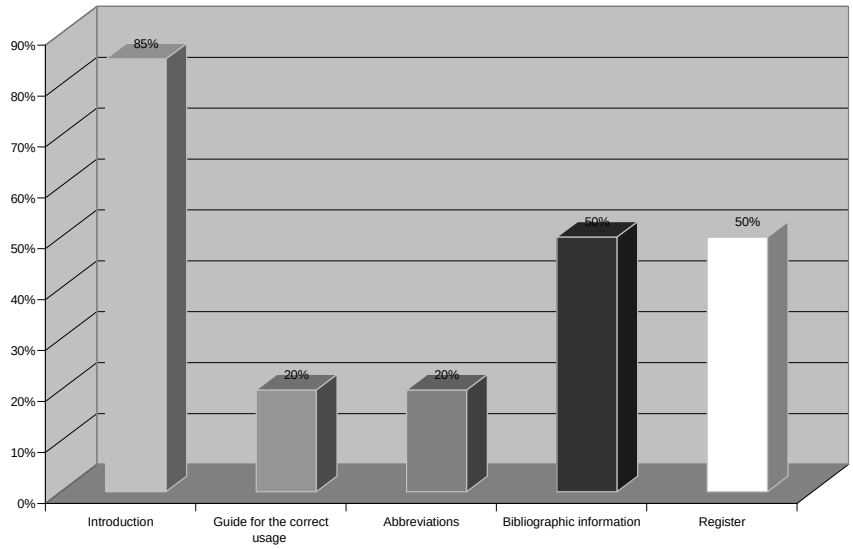


Figure 1. Results from the analyses of the "outside matter"

Figure 1 displays the results of our analyses of the "outside matter". It shows that:

- Many dictionaries (85%) have an introduction (but only 50% have an introduction that follows lexicographic criteria; most of them are made up of only two sentences);
- Only 20% have a usage guide, but it is always very minimal;
- Only 20% have an explanation of the abbreviations (although usually quite simplistic);
- Only 50% have bibliographic information;
- Only 50% have a register or index.

We can conclude that there is a significant difference between written and online dictionaries. Written dictionaries are more academically sound, and they have a more complete macrostructure. Online dictionaries usually do not make use of the most important technological instruments, like querying tools, links, etc. About 70% of the online dictionaries are written dictionaries converted into an online form.

DIL uses as a model the dictionaries of linguistics in their written form, therefore it includes an introduction, an index, a list of abbreviations, a guide for the correct usage of the dictionary, guidelines for the addition of new entries; and bibliographic information at the end of the most important entries.

DIL also attempts to use the specific features of online dictionaries, in that it has querying tools, links (external to other related dictionaries and internal between the entries). Multimedia facilities are also planned.

4. Criteria for the selection of entries and lexicographic basis

The function of the dictionary should always be kept in mind when deciding on the criteria for the selection of entries. These include the user group and the context of use (*cf.* Bergenholtz 1989; Beißenwenger and Körkel 2002; Kromann *et al.* 1984). Even though an LSP dictionary presents fewer problems than a language dictionary, the risk of being too subjective is high (*cf.* Haensch 1991: 2920).

In order to reduce this risk, a lexicographer can resort to:

a) Computational analyses;

b) Analyses of the entry list of other dictionaries of the same or similar type (Barz *et al.* 2005: 88; Bergenholtz 1989: 774);

c) Creation of a corpus of LSP texts and books that belong to the lemmatized field and can be used as a guideline.

DIL used only b) and c) because a) was not suitable for our purpose.

In terms of the subfield of DaF, DIL used two types of sources:

1) General printed dictionaries of linguistics and their list of entries.[7] These sources have been analysed by identifying terms belonging to the lemmatized subfield. In the case of a term pertaining to more disciplines, only the sense belonging to the field of German as a Foreign Language (DaF) was taken into account.

2) LSP dictionaries and glossaries of applied linguistics. Two types have been analysed:

- specific LSP dictionaries or glossaries, such as Balboni (1999) and Homberger (2005);
- small glossaries, that can usually be found at the end of important books and compendiums, such as Balboni (1999) (74 entries), Ciliberti (1994) (40 entries), and Frabboni (1992) (51 entries).

A small corpus of LSP texts was created. General books of applied linguistics and more specific books of DaF[8] were analysed for key words. These were then imported into *Excel* tables that were consulted for the compilation of the final list. The main criteria for the selection of the entries were conceptual relevance and frequency; semantic maps were also of great help.

Lacunae are a typical problem of dictionaries, but in the case of online dictionaries, the gap can easily be filled. In DIL it was decided not to lemmatize the following types of entries:

---

[7] *Cf.* Lewandowski (1994), Glück (2000), Bußmann (2002), Beccaria (2004), and Bußmann (2007).

[8] *Cf.* Bausch *et al.* (1995), Ciliberti (1994), Freddi (1994), Helbig *et al.* (2001), Huneke and Steinig (2005), Ricci Garotti (2004), Roche (2005), Rösler (1994), Storch (1999), Wierlacher (1980).

a) terms that overlap with other disciplines;

b) terms which belonged essentially to the practical teaching;

c) terms that are not real LSP words.

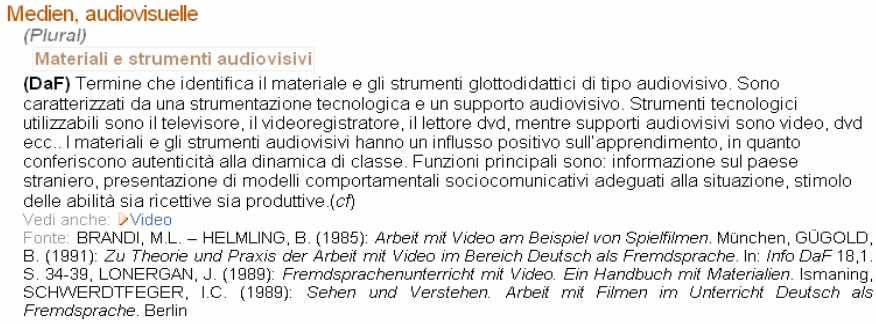
In the case of polysemous words, the different meanings are listed in the explanation of the entry. In the case of synonyms, all of them were listed as entries, but the definition itself is located only under the more frequently used term; the others are linked to this.

5. Conclusion

Two main criteria were used for the selection of entries, *i.e.* conceptual relevance and frequency. A small and ad hoc corpus of texts was thus created consisting of 242 lemmas pertaining to the section of DaF (German as a foreign language). Some sections of the dictionary concerning historical syntax, text linguistics, morphology, language for special purposes and lexicography are currently being prepared, whereas others have been planned (phonology, syntax etc.).

DIL is an encyclopaedic dictionary which, however, wants to give additional information to the potential user on the basis of the needs mentioned in Section 2. So every entry is followed by grammatical information about number[9] and genre, the Italian equivalent or equivalents, the field of linguistics to which the term belongs in an abbreviated form, the definition and explanation, the code corresponding to the author of the entry, related terms and the bibliography. Examples and possible synonyms can also be found.

Figure 2 is an example of an entry taken from the dictionary.


Medien, audiovisuelle

(Plural)

Materiali e strumenti audiovisivi

(DaF) Termine che identifica il materiale e gli strumenti glottodidattici di tipo audiovisivo. Sono caratterizzati da una strumentazione tecnologica e un supporto audiovisivo. Strumenti tecnologici utilizzabili sono il televisore, il videoregistratore, il lettore dvd, mentre supporti audiovisivi sono video, dvd ecc.. I materiali e gli strumenti audiovisivi hanno un influsso positivo sull'apprendimento, in quanto conferiscono autenticità alla dinamica di classe. Funzioni principali sono: informazione sul paese straniero, presentazione di modelli comportamentali sociocomunicativi adeguati alla situazione, stimolo delle abilità sia ricettive sia produttive.(*cf*)

Vedi anche: Video

Fonte: BRANDI, M.L. – HELMLING, B. (1985): *Arbeit mit Video am Beispiel von Spielfilmen*. München, GÜGOLD, B. (1991): *Zu Theorie und Praxis der Arbeit mit Video im Bereich Deutsch als Fremdsprache*. In: *Info DaF* 18,1. S. 34-39, LONERGAN, J. (1989): *Fremdsprachenunterricht mit Video. Ein Handbuch mit Materialien*. Ismaning, SCHWERDTFEGER, I.C. (1989): *Sehen und Verstehen. Arbeit mit Filmen im Unterricht Deutsch als Fremdsprache*. Berlin

Figure 2. Example of an entry of DIL
(http://www.humnet.unipi.it/dott_linggensac/glossword/)

[9] In the case of plural instead of mentioning the article the user will find the word *Plural*. This to avoid misunderstanding because in German the feminine article and the plural article are identical.

References

ABEL, A. (2006). Elektronische Wörterbücher: Neue Wege und Tendenzen. In F. SAN VINCENTE (ed.). *Akten der Tagung "Lessicografia bilingue e Traduzione: metodi, strumenti e approcci attuali" (Forlì, 17.-18.11.2005).* Polimetrica Publisher (Open Access Publications): 35-56.

BALBONI, P.E. (1999). *Dizionario di Glottodidattica*. Perugia: Edizioni Guerra.

BARZ, I., BERGENHOLTZ, H. and KORHONEN, J. (2005). *Schreiben, Verstehen, Übersetzen, Lernen. Zu ein- und zweisprachigen Wörterbüchern mit Deutsch*. Frankfurt a.M.: Peter Lang.

BAUNEBJERG HANSEN, G. (1990). *Artikelstruktur im zweisprachigen Wörterbuch. Überlegungen zur Darbietung von Übersetzungsäquivalenten im Wörterbuchartikel*. Tübingen: Max Niemeyer Verlag (*Lexikographica Series Maior*, 35).

BECCARIA, G.L. (ed.). (2004). *Dizionario di linguistica e di filologia, metrica, retorica*. 3. Auflage. Torino: Einaudi.

BEIßWENGER, M. and KÖRKEL, B. (2002). Die Lemmaselektion im De Gruyter Wörterbuch Deutsch als Fremdsprache. In H.E. Wiegand (ed.). *Perspektiven der pädagogischen Lexikographie des Deutschen. Untersuchungen anhand des "De Gruyter Wörterbuch Deutsch als Fremdsprache"*. Tübingen: Max Niemeyer Verlag: 393-412.

BERGENHOLTZ, H. (1989). Probleme der Selektion im allgemeinen einsprachigen Wörterbuch. In F.J. Hausmann *et al.* (eds). *Wörterbücher: ein internationales Handbuch zur Lexikographie* Band 1. Berlin & New York: De Gruyter: 773-779.

BERGENHOLTZ, H. (1992). Lemmaselektion im zweisprachigen Wörterbuch. In G. Meder and A. Dörner *Worte, Wörter, Wörterbücher. Lexikographische Beiträge zum Essener Linguistischer Kolloquium*. Tübingen: Max Niemeyer Verlag (*Lexikographica. Series Maior* 42).

BERGENHOLTZ, H. and KAUFMANN, U. (1997). Terminography and Lexicography. A Critical Survey of Dictionaries from a Single Specialised Field. In *Hermes*, 18: 91-125.

BUßMANN, H. (2002). *Lexikon der Sprachwissenschaft*. 3. Auflage. Stuttgart: Kröner.

BUßMANN, H. (2007). *Lessico di Linguistica. Traduzione italiana, adattamento e revisione sulla base delle 3° edizione originale, rivista ed ampliata a cura di Paola Cotticelli Kurras.* Alessandria: Edizioni dell'Orso.

CARDONA, G.R. (1988). *Dizionario di linguistica*. Roma: Armando.

CILIBERTI, A. (1994). *Manuale di glottodidattica*. Firenze: La Nuova Italia.

FOSCHI ALBERT, M. AND HEPP, M. (2004). *Zum Projekt: Bausteine zu einem deutsch-italienischen Wörterbuch der Linguistik.* DaF Werkstatt, 4: 43-69.

FRABBONI, F. (1992). *Manuale di didattica generale*. Roma & Bari: Laterza.

GLÜCK, H. (2000). *Metzler Lexikon Sprache*. 2. Auflage. Heidelberg: Metzler.

HAENSCH, G. (1991). Die mehrsprachigen Wörterbücher und ihre Probleme. In F.J. Hausmann, O. Reichmann, H.E. Wiegand and L. Zgusta (eds). *Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch zur Lexikographie. Handbücher zur Sprach- und Kommunikationswissenschaft (HSK 5.3.).* Dritter Teilband. Berlin & New York: De Gruyter: 2909-2937.

HARTMANN, R.R.K. (ed.) (1983). *Lexicography: Principles and practice*. London: Academic Press.

HOMBERGER, D. (2005). *Lexikon Schulpraxis.* München: Schneider Verlag.

KROMANN, H.P., RIIBER, T. and ROSBACH, P. (1984). Überlegungen zu Grundfragen der zweisprachigen Lexikographie. In H.E. Wiegand (ed.). *Studien zur neuhochdeutschen*

Lexikograophie V. Hildesheim & New York & Zürich (Germanistische Linguistik 3-6/84): 159-238.

KÜHN, P. (1983). Sprachkritik und Wörterbuchbenutzung. In H.E. Wiegand (ed.). *Studien zur neuhochdeutschen Lexikographie* III. Hildesheim & New York: Olms (Germanistische Linguistik 1-4/82): 157-177.

LEWANDOWSKI, T. (1994). *Linguistisches Wörterbuch*. 6. Auflage. Heidelberg: Quelle & Meyer.

NIELSEN, S. (1990). Contrastive Description of Dictionaries Covering LSP Communication. *Fachsprache/International Journal of LSP*, 3-4: 129-136.

STORRER, A. and HARRIEHAUSEN, B. (1998). *Hypermedia für Lexikon und Grammatik.* Tübingen: Narr.

WIEGAND, H.E. (1977). Fachsprachen im einsprachigen Wörterbuch. Kritik, Provokationen und praktisch-pragmatische Vorschläge. In H. Schumacher and B. Leuschner (eds). *Linguistik: Beschreibung der Gegenwartssprache. Kongreßberichte der 7. Jahrestagung der Gesellschaft für Angewandte Linguistik. GAL. Trier 1976*. Bd. II. Stuttgart: 39-65.

WIEGAND, H.E. (1998). *Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie*. 1 Teilband. Berlin: De Gruyter.

ZÖFGEN, E. (1991). Bilingual learner's dictionaries. In F.J. Hausmann *et al*. (eds). *Wörterbücher/Dictionaries/Dictionnaries. An International Encyclopedia Lexicography* (III). Berlin & New York: De Gruyter: 2888-2903.

ZÖFGEN, E. (1994). *Lernerwörterbuch in Theorie und Praxis. Ein Beitrag zur Metalexikographie mit besonderer Berücksichtigung des Französischen*. Tübingen: Max Niemeyer Verlag (*Lexikographica Series Maior*, 59).

Acquiring semantics from structured corpora to enrich an existing lexicon

Nuria Gala[1], Véronique Rey[2]
Aix-Marseille Universités

Abstract

Lexical and semantic information is crucial for NLP by and large and for building linguistic resources in particular. However, this kind of information is hard to be obtained be it manually or automatically. The difficulties come from the very nature of the information (what do we mean by 'meaning'? What is 'semantics'? How are 'meanings' combined and put into words?) and from the resources (How are 'meanings' formally represented and displayed?). In this paper we present a methodology for automatically acquiring semantics from structured corpora (machine-readable dictionaries and free encyclopaedias on the web). This information is used to enrich an existing lexical database to build families of words in modern French.

Keywords: automatic acquisition, semantic information, morpho-phonology, lexical database of word families.

1. Introduction

At the age of digital resources, language learners benefit greatly from a variety of available applications. For example, in the case of electronic dictionaries, the learner may look up information concerning a given word: meaning, etymology, spelling, pronunciation, grammar (gender, part of speech), usage (occurrence in a text), etc. In addition, frequency information has become available (Kilgarriff 1996). The learner may thus find out if the word is common or not in the language or in a particular specialized sub-language.

The resources used for e-learning tend to be electronic versions of paper resources: they present the same information, the only difference being the way this information is searched and displayed. Some examples illustrate this point for French: the TLFi[3] (*Trésor de la Langue Française*) and the DAF[4] (*Dictionnaire de l'Académie*

[1] LIF, CNRS UMR 6166, Marseille, France, nuria.gala@lif.univ-mrs.fr
[2] SHADYC, CNRS & EHESS, UMR 8562, Marseille, veronique.rey@univ-provence.fr
[3] http://atilf.atilf.fr/tlf.htm
[4] http://dic.academic.ru/dic.nsf/daf_1835

Française), both available online with exactly the same content as in the paper version.

Lexical databases and networks, built to be used by means of computers (no paper versions), are supposed to go further with regards to the lexical description (due to storage possibilities, the conceptual organisation of the information, etc.). Though a number of projects have arisen, more specifically for multilingual resources (Papillon database[5], EuroWordNet, etc.), the learner only obtains definitions and lists of synonyms when looking for semantic information. It is not feasible for such resources to "navigate" through lexical units sharing semantic components in the same family neither to access a particular lexical unit from a set of ideas carried by the semantic information (as in flexional languages like French the construction of words is based on phonological and semantic continuity, it would be relevant to navigate in a resource by exploiting this implicit practice of speakers).

The goal of the work described here is to propose a method for automatically enriching with semantic information an existing lexical database of modern French, Polymots. Adding semantics to the resource will increase the number of navigational paths, that is, the number of ways to access a given word.

The paper is organized as follows. First, in Section 2, we introduce the notion of morpho-phonological families and we briefly outline the main features of Polymots. In Section 3 we present a method for automatically collecting semantic information from structured corpora and we discuss the notion of continuity and dispersion of meaning within a family of words. Finally, we present our conclusions and future work in Section 4.

2. Morpho-phonological families in modern French

Traditionally, lexical morphology has been diachronic and has focused on the notion of word families on the basis of word form origins (etymology). In morphological synchronic studies, the focus is rather on segmenting words in minimal meaningful units (morphemes). In this approach, the aim is to build models of morphological constructions of the lexicon. However, this task is far from being trivial.

2.1. Morpho-phonology

While in some cases structural analysis is quite straightforward, *i.e.* "bras" (*arm*) and "brassard" (*armband*) clearly share an element, the stem "bras" (*arm*, which is also common in both English translations), in others it is not. Questions such as the following may arise: (1) do "biscuit" (*cookie*) and "cuire" (*to bake*) belong to the same family?; (2) can we group into the same family "*confiture*" (jam) and "défaite"

[5] http://www.papillon-dictionary.org

(*defeat*)? In the first case, we recognize the past participle form of the French verb to bake ("cuit", *baked*); in the second, we perceive a link between two inflected forms (the simple past form ("fit", *did*) and the past participle ("fait", *done*)) of the same verb "faire" (*to do*). In neither of these cases, do we perceive the same form, but rather one of many possible ones.

In consequence, and following Kiparsky (1982), we assume that the process of word construction implies phonological transformations, that is, vocalic and consonantic alternations. As the process takes place within the lexeme, we can talk about morpho-phonological processes. However, the alternations are not systematic. To give an example, while many words ending in /o/ alternate with /el/ ("ciseau" *chisel*, "ciseler" *to chisel*; "château" *castle*, "châtelain" *manor*; "appeau" *decoy*, "appel" *call*) some do not ("fourreau" *sleeve*, "berceau" *cradle*, "gâteau" *cake*). In order to take into account this morphophonemic process, we have annotated manually the words of our word family corpus.

2.2. Semantics

Segmenting words into morphemes raises interesting questions concerning meaning, as well as morphemes and word families as a whole.

2.2.1. Morphemes

Some words in French have been created with meaningless morphemes (we call them "opaque stems"). To give an example, the word "tri-maran" has been built following "cata-maran", although "maran" is a non-Latin meaningless stem. Other examples are "panta-lon" and "panta-court" which share the stem "panta".

In various cases, it is possible to identify a common stem in a family which does not have a meaning as a single word in modern French (despite its Latin origin). For example, "duct" is not a lemma (lexical unit); however, it is part of "con-duct-eur" (*driver*), "pro-duct-eur" (*producer*), "intro-duct-ion", "sé-duct-eur" (*seducer*), etc.

2.2.2. Word Families

The point that we would like to make in this paper is that words in families may have common meanings. As explained in Section 1.1., words are grouped into families on the basis of a common morpho-phonological stem. Therefore, in some families the segmentation entails the problem of lexical variation. To illustrate this idea: if words in the previous family share a stem ("duct"), the question that can be asked is the following: What is the nature of the (semantic) link between all the words in this family ? This question is implied by our hypothesis that all the words in a family share not only a morpho-phonological stem, but also have some kind of meaning, hence, are semantically coherent. In some cases, the common meaning is quite straightforward: "terre" (*globe/earth*), "territoire" (*territory*) and "terrasse" (*terrace*) share the notion of

area and *surface*; “gluant” (*viscous*), “glutineux” (*sticky*), “agglutiner” (*aglutinate*) may be said to share the notions of *sticky, adhesive, viscous*, etc.

Clearly, the degree of semantic cohesion in a family is variable. In some cases it is transparent, in other cases it is much harder to grasp and, as a consequence, it entails substantial conjectures about the semantic continuum in a family. That is the point we wanted to investigate by automatically acquiring semantic information from structured corpora (*cf.* Section 2). Before describing our method, we briefly outline the features of the lexical database developed for French words organized into families.

2.3. Polymots

After manual segmentation of 20,000 French words we grouped the words to build a lexical database of about 2,000 families (Gala and Rey 2008). Subsequent morphological analysis revealed that one third of them are “opaque stems” and the other two thirds “transparent stems”. Transparent stems are meaningful words in French, *e.g.* “terre” (*earth/globe*), “glue” (*glu*), and “boule” (*ball*).

Constructional morphology is very common in French and other Romance languages, the average of words in a family being about ten lexical items. However, productivity varies with the families. Unlike families composed of one or two members (*e.g.* “chaise” (*chair*) is the only lexical unit of its family, “choi” being the common stem of “choix” (*choice*) and “choisir” (*to choose*)), some stems can be found in various families containing up to seventy or eighty lexical items: “mue/mut” in “commuter” (*to commute*), “immuable” (*immutable*), “mutuel” (*mutual*), “remuer” (*to shake*), etc..

3. Acquisition of semantic information from structured corpora

At present, Polymots displays only morphological information which is used by speech therapists to help patients presenting certain diseases (*e.g.* dyslexia, Alzheimer) improve vocabulary learning. The need for semantic information in this context is twofold: understand unknown lexical items and be able to access them.

This being so, we consider that the learner would probably better understand the meaning of a lexical item by grasping the semantic links connecting words with other words of the same family (hence the meaning of unknown words like *gluey* or *glueball* would easier be grasped if we mentioned *glue*, the ‘baseword’, as it shares with the target words notions like *stickiness, adhesion, viscousity*). As shown by Zock and Schwab (2008), adding semantics will also help the language producer (speaker/writer) find the word s/he is looking for by providing some kind of conceptual input (for example, *file, key, path, fast* yielding *shortcut*).

3.1. Related work concerning semantic acquisition

Automatically acquiring information from digital corpora is one of the well studied tasks in natural language processing (NLP). However, the construction and enrichment of electronic resources from corpora is far from being trivial, be it only for the sparseness of available data. Leaving aside manually built resources (extremely time-consuming), a number of studies have been carried out using the web (Grefenstette 2007).

Other approaches have used existing resources such as dictionaries, synonym lists, ontologies, etc. as information is well structured, available and easily exploitable. Dictionary definitions are used for many applications: creation of lexical networks (Ide and Véronis 1990), building of an example database for semantic disambiguation (Brun *et al.* 2001), acquisition of conceptual links between words (L'Homme 2003), building of lexical graphs (Gaume *et al.* 2007), etc.

3.2. Structured corpora

Concerning word families, reliance on structured corpora is crucial for being able to enrich the words being part of a given family. The idea here is to collect information from different structured resources to build a list of semantic units describing each word.

One of the difficulties in using structured corpora is their availability. For reasons of copyright, most of the dictionaries with online consultation are not available in an exploitable format (text format or, better, XML format). For example, this is the case for one of the main French dictionaries, the TLFi (Trésor de la Langue Française Informatisé).

In order to diversify our sources, we used the following lexicographic and encyclopaedic resources:

- Hachette Multimédia dictionary (in XML format)
- Wiktionnaire[6] (French version of Wiktionary)
- French Wikipedia[7]

Our aim was to collect the meaningful words present in the dictionaries' definitions and in the introductory paragraph of Wikipedia. Using our list of 20,000 words we collected the different entries by relying on the above mentioned sources.

In the case of Wikis, we retrieved the required web pages, corresponding to our list of words. From Wikipedia, we only retrieved the introductory paragraph of each article. We skipped other encyclopaedic information as we considered it irrelevant for our purpose.

[6] http://fr.wiktionary.org/wiki/
[7] http://fr.wikipedia.org/wiki/

After removing the HTML tags, we obtained text files as shown in Figures 1 and 2 for the word “vache” (*cow*), the example used to illustrate our method.

vache féminin

(Zoologie) Mammifère domestique ruminant, généralement porteur de cornes sur le front, appartenant à l’espèce *Bos taurus* de la famille des bovidés.

Femelle de cette espèce.

Figure 1. Sample obtained from Wiktionnaire

Vache (Brune Suisse ou Brune des Alpes) vue sous la Fuorcla Sesvenna dans l'Engadine, en Suisse.

La vache est la femelle d'un mammifère domestique ruminant, généralement porteur de cornes sur le front, appartenant à l'espèce Bos taurus de la famille des bovidés. C'est la femelle du taureau. Une génisse est une vache qui n'a pas vêlé.

Le poids moyen d'une vache adulte varie en fonction de la race de 500 à 900 kg.

Le mot vache vient probablement du sanscrit Vaça désignant une génisse qui vêle pour la première fois.

Figure 2. Sample obtained from Wikipédia

3.3. Methodology to obtain semantic units

We tested the hypothesis that each definition contains significant lexical items to characterize semantically the words of our families. To this end we grouped the corpora by headwords and extracted the meaningful words after having removed stopwords (prepositions, articles), frequent adverbs, conjunctions and various nouns such as ‘verb’, ‘synonym’, ‘example’, ‘Latin’, etc.. Since we are interested in lemmata rather than its possible flexional variants or forms, we used the Treetagger[8] to strip off the irrelevant morphological information.

For each entry, the Treetagger’s output was transformed into a vector of words as shown in Figure 3:

vache	femelle bovin
vache	manoeuvrer attaque sournois
vache	peau cuir animal
vache	récipient plier toile plastique analogue utiliser campeur

Figure 3. Meaningful lemmatized words of Hachette’s definitions

[8] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Note that definitions are kept with their headwords, as the order of the lexical elements is needed in order to reveal the word's relative importance.

Errors of the Treetagger may have an impact, especially with regard to noun/verb ambiguities, and this may be a weakness of this approach. In the following examples, nouns have been tagged as verbs: "manoeuvrer" (to maneuver), instead of "manoeuvre" (maneuver); "corner" (to corner, to honk) instead of "corne" (horn); "membrer" (~to equip) instead of "membre" (member), etc. We are currently trying to fix this problem by improving the training with the tagger.

3.4. Weighting semantic units

Once the meaningful words, henceforth semantic units (SU), are lemmatized, we try to evaluate their relative "importance", that is, we try to determine empirically their relation with respect to the definition's headword. As our corpus is made up of three different sources, we hypothesized that significant units will appear in most of the definitions and generally at the beginning. For example, in the case of "vache" (cow), "femelle" (female) is more important than "génisse" (heifer), as "femelle" appears twice (at the beginning of two definitions), while "génisse" appears in the middle of a single definition (or more precisely, the sentence of an entry of an encyclopedia). This being so, the two words present different importance, a fact that needs to be taken into account.

To do so we attributed a weight to each word of the definition, taking into account the distance between each one of them and the headword. Formally speaking, for each headword w, we build a list $\alpha(w)$ composed of its semantic units u (bag of words). Given a semantic unit u_i in this list, we calculate its weight $\omega(u_i)$ on the basis of the distance of u_i and w by taking into account the total number of words n in each definition (with $0 \leq i < n$):

$$\forall\, u_i \in \alpha(w),\ \omega(u_i) = 1 - i/n$$

The relevance of a semantic unit u_i decreases proportionally to its distance to the headword w. If there are four words in the definition, the first one will be assigned 1, the second $1 - 2 / 4 = 0.5$, the third $1 - 3 / 4 = 0.75$ etc. (Gala *et al.* 2009).

A final adjustment is necessary for words appearing more than once within the definitions for a given headword. In this case we sum their weights and harmonize this result by adding all weights (bringing them to a maximal $\omega = 1$ with ω being > 0). To give an example: "femelle" (female) appears twice with $\omega = .$ ı; "mammifère" (mammal) appears twice with $\omega = 0.94$ and $\omega = 0.93$ which yields the final weight of 0.58. Figure 4 shows the results for the entry "vache" (cow).

As "cow" is a polysemic word in French, the semantic units within the vector show the following meanings: "mammifère" (mammal), "cuir" (cowhide), "sournois" (rotten, mean), "toile" (tent). A vector obtained with this method may contain synonyms (embrace and enclose, swallow and go down), hyperonyms (mammal and heifer, alarm

and device) as well as syntagmatic or thematic links (alarm and enemy, swallow and throat), etc.

[femelle 1.00] [mammifère 0.58] [domestique 0.54] [ruminer 0.50] [porteur 0.45] [espèce 0.43] [corner 0.41] [front 0.37] [appartenir 0.32] [adulte 0.31] [manoeuvrer 0.31] [peau 0.31] [récipient 0.31] [vêler 0.31] [zoologie 0.31] [plier 0.27] [varier 0.25] [bos 0.23] [toile 0.22] [attaque 0.21] [cuir 0.21] [poids 0.21] [taurus 0.19] [fonction 0.19] [plastique 0.18] [bovin 0.16] [famille 0.15] [analogue 0.13] [race 0.13] [animal 0.10] [moyen 0.10] [sournois 0.10] [bovidés 0.10] [utiliser 0.09] [mot 0.06] [campeur 0.04] [taureau 0.04] [génisse 0.02]

Figure 4. Semantic units obtained after weighting

4. Continuity *vs* dispersion of meaning in a family

Morpho-phonological families are based on two criteria, phonological and semantical. A closer look at the semantic units of the vectors has led us to make a distinction between terms, some expressing semantic continuity others semantic dispersion.

4.1. Semantic continuity

Semantic continuity is the property of families sharing semantic units. To be more precise, some semantic units are kept within the family and, in most cases, a recurrent word (the transparent stem) is present in the vectors of all the family members. Figure 5 illustrates such families:

terre (*earth/globe*)	surface	bras (*arm*)	membre (*member*)
territoire (*territory*)	surface	brassard (*armband*)	bras (*arm*)
terrasse (*terrace*)	surface	embrasser (*to embrace*)	bras (*arm*)
		bracelet (*bangle*)	bras (*arm*)

Figure 5. Semantic continuity

There is an explicit continuity of meaning among the words expressed via a semantic unit shared by all family members.

4.2. Semantic dispersion

Semantic dispersion is the property of families where a common semantic unit is present only in some of the words of the family. In such a case, only one, or few semantic units are shared (between a word in the family and the ‘headword’ or stem) as shown in Figure 6.

fil (thr*ead*)	long, continuité, fin (*long*, *continuity*, *thin*)
défilé (*parade*)	long, continuité
profil (*profile*)	fin
val (*glen*)	aire, descente (*area*, *downhill*)
vallée (*valley*)	aire
avaler (*swallow*)	descente

Figure 6. Semantic dispersion

Recurrent semantic units characterize words in families with meaning being distributed. Even though these meaningful units may be different, they are all to be found within the vector characterizing the 'headword' corresponding to the stem. In cases where the stem corresponds to a non existing word in modern French ("opaque stems", *cf.* Section 1.2.1), a number of common semantic units are to be found within the family, *i.e.* the notion of dead and dangerous being part of the family with the opaque stem "cid" ("accident", "suicide", "incident", "acide", etc.).

5. Conclusion and future work

We presented a method for automatically enriching an existing lexical database of French with semantic information. The initial idea was to create a lexical resource of word families based on word constructions and to take a morphophonemic approach. As we assumed that words within a given family would share certain semantic features, we gathered this semantic information from structured corpora to empirically validate our hypothesis. An analysis of our results showed that there are two types of families, depending on the dispersion or the continuity of meaning between the words in a family.

The characteristics of word families in Polymots offer a new perspective concerning the study of words as it is based on phonological stems and actual language usage (synchrony) rather than traditional lemmas looked at from a diachronical perspective. Hence this resource is based on a new lexicographical approach, offering new possibilities for learning French or accessing words, by taking phonological and semantic information into account.

References

BRUN C., JACQUEMIN B. and SEGOND F. (2001). Exploitation de dictionnaires électroniques pour la désambiguïsation sémantique. *Revue TAL, Traitement Automatique des Langues,* volume 42(3): 667-691.

GALA N. and REY V. (2008). Polymots : une base de données de constructions dérivationnelles en français à partir de radicaux phonologiques. In *Actes de TALN 08: Traitement Automatique des Langues Naturelles,* Avignon, juin 2008.

GALA N., REY V. and TICHIT, L. (2009). Dispersion sémantique dans des familles morpho-phonologiques: éléments théoriques et empiriques. In *Actes de TALN 09: Traitement Automatique des Langues Naturelles*, Senlis, Juin 2009.

GAUME B., DUVIGNAU K. and VANHOVE M. (2007). Semantic associations and confluences in paradigmatic networks. In: M. Vanhove (ed.). *Typologie des rapprochements sémantique.* John Benjamins Publishing Company.

GREFENSTETTE, G. (2007). Conquering Language: Using NLP on a Massive Scale to Build High Dimensional Language Models from the Web. In *Computational Linguistics and Intelligent Text Processing*. Springer Berlin/Heidelberg, 35-49.

IDE N. and VÉRONIS J. (1990). Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries. In *13th International Conference on Computational Linguistics COLING'90,* vol. 2: 389-394. Helsinki.

KIPARSKY P. (1982). From cyclic Phonology to lexical Phonology. In H.V.D. Hulst and N. Smith (eds). *The structure of Phonological Representations,* Dordrecht: 131-175.

L'HOMME M.-Cl. (2003). Acquisition de liens conceptuels entre termes à partir de leur définition. *Cahiers de lexicologie*, 83(2): 19-34.

KILGARRIFF, A. (1996). Putting frequencies in a dictionary. *International Journal of Lexicography*, 10(2): 135-155.

ZOCK, M. and SCHWAB, D. (2008). Lexical Access based on Underspecified Input. In *Proceedings of COGALEX COLING Workshop endorsed by SIGLEX, Cognitive Aspects of the Lexicon: Enhancing the Structure, Indexes and Entry Points of Electronic Dictionaries*, Manchester, UK: 9-17.

Customising a general EAP dictionary to meet learner needs

Sylviane Granger[1], Magali Paquot[1]
Université catholique de Louvain, Louvain-la-Neuve

Abstract

In this article, we describe the *Louvain EAP Dictionary* (*LEAD*), a web-based English for Academic Purposes dictionary-cum-writing aid for non-native writers. The *LEAD* dictionary contains a rich description of non-technical words that express key functions in academic discourse (such as contrast, exemplification or cause and effect), with particular focus on their phraseology (collocations and recurrent phrases). The dictionary allows for both semasiological and onomasiological access. Its main originality is its customisability: the content is automatically adapted to users' needs in terms of discipline and mother tongue background.

Keywords: academic vocabulary, English for Academic Purposes, English for Specific Purposes, discipline-specific corpora, phraseology, EFL, learner corpus.

1. Introduction

As English is incontestably the dominant language in academia, acquiring good English academic skills is mandatory for the large proportion of users for whom English is a non-native language. More and more university students have to write term papers, reports, or their MA/PhD dissertations in English. The number of master and doctoral programmes taught in English has increased dramatically over the last decade. For researchers, the stakes are even higher as inappropriate language use is a major factor in the rejection of articles submitted to international journals by non-native writers (*cf.* Mungra and Weber 2010). And crucially, a highly revealing correlation has been found between national output of international scientific publications and national English proficiency level (Man *et al.* 2004).

There is obviously an urgent need to design tools that meet the needs of non-native speakers, be they English as a Foreign/Second Language (EFL/ESL) students or young researchers. In order to help them master the vocabulary needed in academic settings, a number of word lists have been compiled. The *Academic Word List* (AWL) (Coxhead 2000), for example, consists of 570 word families which have wide range

[1] Centre for English Corpus Linguistics, Université catholique de Louvain, {sylviane.granger, magali.paquot}@uclouvain.be

and reasonable frequency of occurrence in a large corpus of academic texts, but are not among the 2,000 most frequent English words. Hyland and Tse (2007: 238), however, note considerable differences in the use of academic vocabulary across disciplines and call into question the very existence of a general academic vocabulary. They argue that "all disciplines shape words for their own uses" (ibid: 240), as demonstrated by their clear preferences for particular meanings and collocations. The noun *strategy*, for example, has different preferred associations across disciplines (*e.g. marketing strategy* in business, *learning strategy* in applied linguistics and *coping strategy* in sociology).

While this register variability is a source of difficulty for native and non-native writers alike, non-native writers have a harder time as they also tend to transfer into English words and phrases that are typical of academic discourse in their first language (L1). This tendency induces a wide range of difficulties: semantic errors (*e.g.* French learners' often wrongly use '*indeed* in the sense of '*en effet*'), lexico-grammatical errors (*e.g.* French learners will write '*discuss about*' mapped on French '*discuter de*'), awkward collocations (*e.g. put forward a conclusion; a conclusion must be gathered*), and rhetorical or stylistic infelicities (*e.g.* French learners tend to overuse sequences with 1st person plural imperative – '*let us examine*', '*let us take the example of*' – which reflect different rhetorical conventions in French academic writing) (see Gilquin *et al.* 2007; Paquot 2008, 2010; Granger and Paquot 2009b).

In this article, we describe the rationale behind an electronic tool which aims to meet some of the difficulties that beset EFL writers in academic settings. Section 2 describes academic vocabulary and focuses more particularly on the difficulties it poses to non-native writers. Section 3 discusses a key lexicographic development afforded by the electronic medium, *i.e.* dictionary customisation. Section 4 introduces the *Louvain EAP Dictionary* (*LEAD*), a web-based EAP dictionary-cum-writing aid tool for non-native writers and Section 5 concludes.

2. Academic vocabulary, phraseology and learner writing

As demonstrated by Howarth's (1996/1998) study of verb + noun collocations in a corpus of social science texts, a large proportion of non-technical collocations in academic writing consist of a verb in a figurative sense and an abstract noun (*e.g. adopt + approach, reach + conclusion, obtain + result*). Howarth suggests that these collocations are an essential part of the procedural vocabulary of academic discourse. The author further argues that it is "not idioms that learners need for effective communication" (Howarth 1996: 156), at least in academic settings. Learners need the lexical means that will allow them to conform to "the native stylistic norms for a particular register", which "entails not only making appropriate grammatical and lexical choices but also selecting conventional [multi-word units] to an appropriate extent" (Howarth 1998: 186). Recent studies have shown that the highly

conventionalised nature of academic discourse stems largely from 'lexical extensions' of a set of academic words such as *conclusion, issue, claim* or *argue*. These words acquire their organisational or rhetorical function in specific word combinations that are essentially semantically and syntactically compositional (*e.g. as discussed below, an example of … is …, the aim of this study is to..., it has been suggested, final outcome, direct result*) (*e.g.* Curado Fuentes 2001; Pecman 2008; Siepmann 2005). Most of these studies have also highlighted the extent to which there is commonality across academic genres and disciplines and thereby brought support to Gledhill's view that "there is a shared scientific voice or 'phraseological accent' which leads much technical writing to polarise around a number of stock phrases" (Gledhill 2000: 204).

Learner corpus research has shown that there is little variation in the way EFL learners organise their papers and that they make scant use of lexico-grammatical patterns typical of academic discourse. In her study of verb + noun combinations in German learner writing, Nesselhauf (2005) notes that "the unavailability of pragmatic chunks for the learners (…) appears to be the underlying reason for a number of deviant collocations which are used to structure the body of the essay, (to introduce examples, for instance)" (2005: 141). Chunks such as *Only have a look at, If you have a look at, Let us have a look at, A first argument I want to name for this* are good illustrations of this kind of pragmatic failure. De Cock's (2003) study of prefabs in learner writing illustrates another aspect of stylistic deficiency, viz. learners' tendency to overuse a whole set of informal word sequences such as *and so, I think* and *there are a lot of* that confers a speech-like quality to their writing. She shows that learners are generally unaware of "the more common, less salient and frequently used L2 multi-word building blocks" (De Cock 2003: 65).

Learners' use of phraseological patterns is also characterised by erroneous collocations and first language influence. Nesselhauf (2005), for example, shows that the most frequent types of error in verb + noun combinations produced by German EFL learners involve the erroneous choice of verb and in 56% of cases, are likely to result from transfer from the learner's L1. Among the nouns that are most often used with deviant verbs are *action, aim, attitude, problem, question, statement, step* and *conclusion*, which fulfil key rhetorical functions in academic discourse. De Cock (2003) also showed that French learners (1) misuse English sequences that have a French congruent form which may be used differently, *e.g. on the contrary*/*au contraire*; (2) underuse multi-word units which have no literal L1 counterpart, *e.g. sort of;* and (3) use L1-induced idiosyncratic combinations, *e.g. according to me*.

3. Dictionary customisation

Dictionaries have traditionally been designed as "one-size-fits-all package[s]" (Rundell 2007: 50). Learners' dictionaries, in particular, target a generic learner and claim to cater for their supposedly similar needs. One of the reasons behind this format

is purely economic: it stems from the wish to reach the widest possible market with one single product and thereby maximise profits. This position is no longer defensible today. Recent research has illustrated the clear necessity to adapt dictionaries to users' needs and technological advances have simultaneously put this development within the reach of dictionary producers. For many specialists, customisation is one of the main challenges of present-day lexicography. Tarp (2009a: 25), for example, points out that "lexicographic needs are not abstract needs, but are always related to specific types of users who find themselves in a specific type of social situation". Put differently, "users in general never need information in general" (Tarp 2009b: 46). One way of implementing customisation is to replace the static data in electronic dictionaries "by articles containing dynamic data which are, so to say, unique for each search related to a specific type of user in a specific type of user situation" (Tarp (2009a: 29). It must be admitted, however, that there have been but few concrete achievements to date in spite of the fact that the idea of dictionary customisation has been around for quite some time (*cf.* Atkins 2002; Varantola 2002; De Schryver 2003). As noted by Sobkowiak (2002), only superficial elements of customisation have been integrated:

> [O]nly the rather superficial customizing options are offered, such as, for example: (a) ignoring certain elements of the entry (micro)structure for screen display (*e.g.* phonetic transcription) or in full-text search (*e.g.* example sentences), (b) hiding certain word categories (*e.g.* compounds), (c) changing font size, style and colours, (d) manipulating toolbars, and the like. (Sobkowiak 2002)

As the difficulties posed by academic English have proved to vary according to users' profile and context of use, in particular their mother tongue background and the discipline they write in, it is a field that would greatly benefit from customisation. It was this that prompted us to embark on the development of a customisable EAP dictionary which will be briefly outlined in the following section.

4. The Louvain EAP Dictionary

The *Louvain EAP Dictionary* (*LEAD*) is a corpus-based tool: it is based on the analysis of c. 900 academic words and phrases in a large corpus of academic texts (*i.e.* the academic component of the British National Corpus as well as a number of home-made discipline-specific corpora) and EFL learner corpora representing a wide range of L1 populations. As shown in Figure 1, the dictionary contains a rich description of academic words, with particular focus on their phraseology (collocations and recurrent phrases). Its main originality is its customisability: the content is automatically adapted to users' needs in terms of **discipline** and **mother tongue background**. The dictionary relies on a relational MySQL database, the technical characteristics of which make it possible to exploit linguistic information as a 'multifunctional lexicographical database', *i.e.* a "modularly designed dictionary database targeting several kinds of users in many different user situations" (Pajzs 2009: 326).

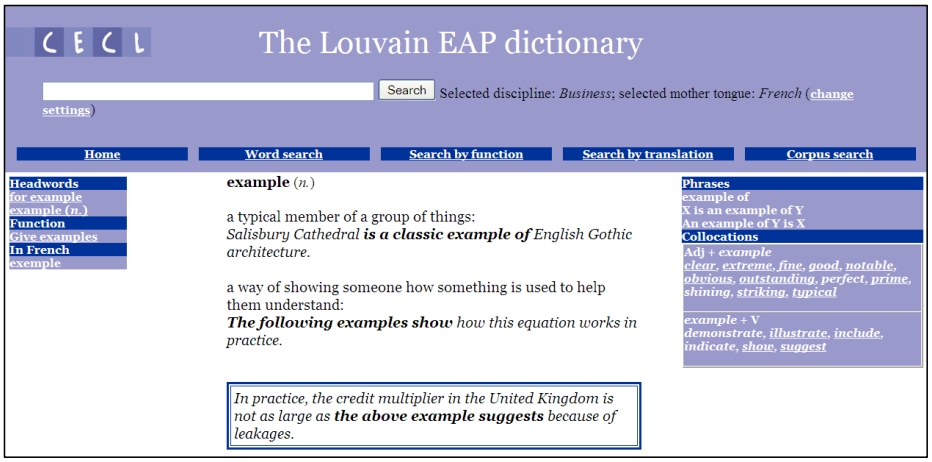


Figure 1. The Louvain EAP dictionary

A key feature of the *LEAD* is that it makes full use of the capabilities afforded by the electronic medium in terms of multiplicity of access modes (Sobkowiak 2002; Tarp 2009). The dictionary can be used as both a **semasiological** dictionary (from lexeme to meaning) and an **onomasiological** dictionary (from meaning/concept to lexeme) via a list of typical rhetorical or organisational functions in academic discourse (*cf.* Pecman 2008). It is also a **semi-bilingual dictionary** (*cf.* Laufer and Levitzky-Aviad 2006) as users who have selected a particular mother tongue background can search lexical entries via their translations into that language.

Before using the dictionary, users select a domain (currently business, medicine, linguistics, or general EAP for users working in other disciplines) and specify their L1 background (currently French) (*cf.* Figure 2). This stage conforms to Tarp's (2009b: 48) suggestion "to prepare a preliminary interactive phase where the lexicographic tool helps the users to identify and specify their concrete needs before being guided to the corresponding data". Discipline-oriented customisation is currently being implemented in the selection of **examples** of collocations and phrases. The characteristics of good dictionary examples have been clearly identified by Atkins and Rundell (2008: 458): they should be (1) natural and typical, (2) informative, and (3) intelligible. However, these are not intrinsic properties and they need to be customised to the type of dictionary and the needs of its users. In the *LEAD*, the collocation *cause* + *distress* is illustrated by example (1) when the user has selected business as the target discipline and example (2) when medicine is the target, thereby adhering to Moon's (2008: 333) recommendation that particular attention be paid to "the function of phraseological information in relation to the needs and interests of the target users".

1. Rivals may not be able to bear initial losses, which would *cause* financial *distress* rather than lead to balanced growth.

2. Severe hypoglycaemic attacks *cause distress* for diabetics and their families.

As shown in Figure 1, clicking on a specific collocate or phrase (*e.g. suggest*) displays a discipline-specific example of the phraseological pattern (*as the above example suggests*) in a box at the bottom of the lexical entry.

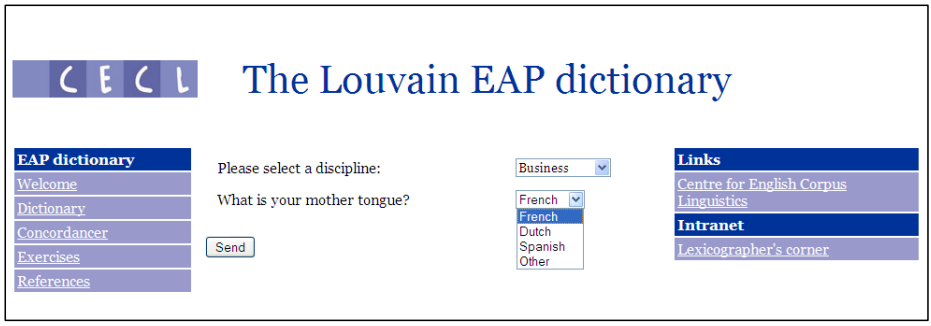


Figure 2. Customising the Louvain EAP dictionary

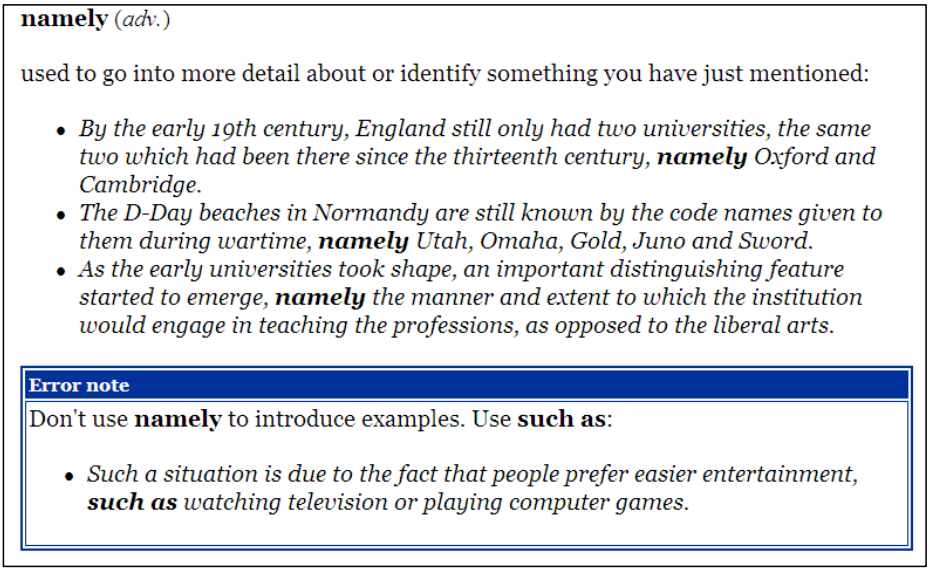


Figure 3. An example of a generic error note

One of the purposes of L1-background identification is to give **feedback** on errors and problems that a specific L1 population typically encounters. When the dictionary is used as a semi-bilingual dictionary, warnings about common translation mistakes are also included, such as the erroneous translation of the French 'prétendre' by its false friend 'pretend' in English. We are currently focusing on French as an L1 background but are planning to include more languages in the future. To create both the generic usage notes and the L1-specific notes, we make use of the *International Corpus of*

Learner English (Granger *et al.* 2009) as well as of the *Varieties of English for Specific Purposes dAtabase* (*VESPA*), a new learner corpus, currently being developed at the Centre for English Corpus Linguistics in collaboration with several international partners. The corpus contains L2 texts from a wide range of L1 backgrounds (currently French, Spanish, Swedish, and Polish), disciplines (linguistics, business, engineering, sociology, etc), genres (papers, reports, MA dissertations) and degrees of writer expertise in academic settings (from first-year students to PhD students) (see http://cecl.fltr.ucl.ac.be/VESPA.html for further details). Errors and difficulties found in the writing of a wide range of learner populations are dealt with in generic error notes that are displayed irrespective of the L1 background selected by the user (*cf.* Figure 3). Errors found exclusively in the writing of French learners are described in notes that only show up if French is selected as L1 background. Thus, the lexical entry for "according to" includes an error note that draws French users' attention to the erroneous translation of French "selon moi" by English "according to me".

Onomasiological access to the dictionary is via a list of 18 rhetorical functions that we have identified as being particularly prominent in academic discourse, *e.g.* comparing and contrasting, expressing cause and effect, introducing a concession (Figure 4).

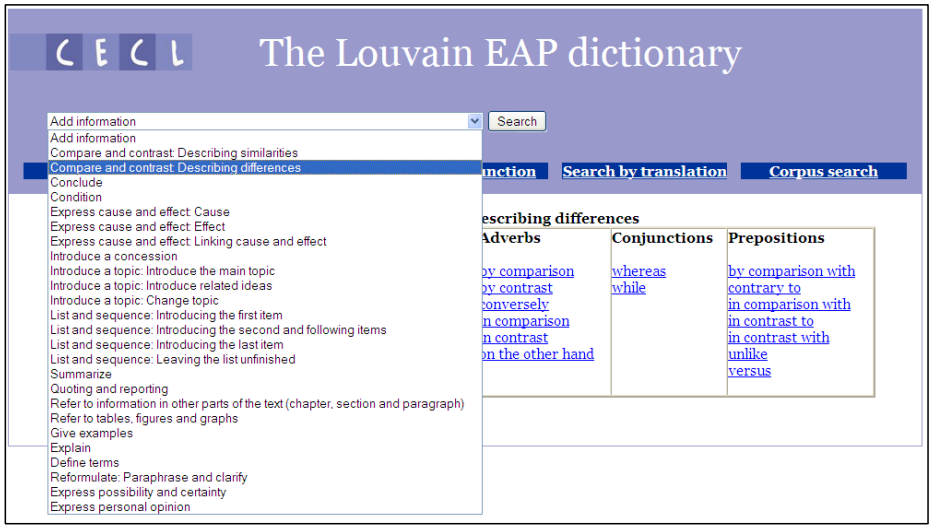


Figure 4. Onomasiological access

Selecting one of these functions provides the user with a list of lexical items, categorised according to part-of-speech (nouns, verbs, adjectives, adverbs, conjunctions and prepositions), typically used to serve this function in academic texts. One of the main advantages of this access mode is that it suggests alternatives and thereby helps users enlarge their academic repertoire. Words are currently sorted alphabetically but they could be sorted by frequency of occurrence in the discipline-

specific corpora. Each word is clickable and users can get access to its full lexical entry.

The *LEAD* dictionary is designed as an integrated tool where the actual dictionary part is linked up to other language resources (in particular, discipline-specific corpora and a corpus handling tool). In the future, we want to turn the *LEAD* into a dictionary-cum-CALL resource (Abel this volume) by adding exercises targeting learners' attested difficulties.

5. Conclusion

Recent research on written academic skills has considerably improved our understanding of the challenges faced by non-native speakers when they write academic texts in English. In particular, it has uncovered the role played by non-technical academic words to express key academic functions such as contrasting or reporting. Corpus-based analyses have demonstrated a high degree of commonality in the use of these words by expert writers from different disciplines but have also highlighted a number of discipline-specific patterns that need to be described. At the same time, learner corpus research has identified the particular types of difficulty that these words pose to non-native writers and demonstrated the important role played by transfer from the learner's mother tongue. Parallel to these findings, recent research has put needs analysis at the heart of both EAP course design and teaching (*e.g.* Jordan 1997; Hyland 2002) and lexicography (Tarp 2008). The *Louvain EAP Dictionary* is an attempt to implement these findings in a customisable web-based tool. While the current version of the tool is restricted to some disciplines and mother tongue backgrounds, its flexible architecture allows for further customisation (other L1 background populations, other disciplines, other languages). The dictionary is currently a stand-alone product but it could – and ideally should – be integrated into a general dictionary and/or a suite of teaching and learning tools.

Acknowledgements

We gratefully acknowledge the support of the Fonds National de la Recherche Scientifique – FNRS, which funded this research within the framework of a project entitled "Lexicography and phraseology: onomasiological and semasiological approach to English for Academic Purposes" (FRFC 2.4501.08).

References

ATKINS, B.T.S. (2002). Bilingual dictionaries – Past, present and future. In M.-H. Corréard (ed.). *Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins*. Euralex: 1-29.

ATKINS, S. and RUNDELL, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

COXHEAD, A. (2000). A new Academic Word List. *TESOL Quarterly*, 34(2): 213-238.

CURADO FUENTES, A. (2001). Lexical behaviour in academic and technical corpora: Implications for ESP development. *Language Learning & Technology*, 5(3): 106-129.

DE COCK, S. (2003). *Recurrent sequences of words in native speaker and advanced learner spoken and written English: a corpus-driven approach*. Unpublished PhD thesis. Louvain-la-Neuve: Université catholique de Louvain.

DE SCHRYVER, G.-M. (2003). Lexicographers' dreams in the electronic-dictionary age. *International Journal of Lexicography*, 16(2): 143-199.

GILQUIN, G., GRANGER, S. and PAQUOT, M. (2007). Learner corpora: the missing link in EAP pedagogy. In P. Thompson (ed.). *Corpus-based EAP Pedagogy*. Special issue of *Journal of English for Academic Purposes*, 6(4): 319-335.

GLEDHILL C. (2000). *Collocations in Science Writing*. Language in Performance 22. Tuebingen: Gunter Narr Verlag.

GRANGER, S., DAGNEAUX, E., MEUNIER, F. and PAQUOT, M. (2009). *The International Corpus of Learner English. Handbook and CD-ROM. Version 2.* Louvain-la-Neuve: Presses universitaires de Louvain. Available from http://www.i6doc.com.

GRANGER, S. and PAQUOT, M. (2008). From dictionary to phrasebook? In E. Bernal and J. DeCesaris (eds). In *Proceedings of the XIII EURALEX International Congress.* Barcelona, 15-19 July 2008: 1345-1355.

GRANGER, S. and PAQUOT, M. (2009a). In search of General Academic English: A corpus-driven study. *Proceedings of the International Conference on L.S.P: 'Options and Practices of LSP Practitioners'*, 7-8 February 2009, Heraklion. Available from http://cecl.fltr.ucl.ac.be/publications.html.

GRANGER, S. and PAQUOT, M. (2009b). Lexical verbs in academic discourse: A corpus-driven study of learner use. In M. Charles, S. Hunston and D. Pecorari (eds). *At the Interface of Corpus and Discourse: Analysing Academic Discourses*. London: Continuum: 193-214.

HOWARTH, P. (1996). *Phraseology in English Academic Writing: Some Implications for language learning and dictionary making*. Tübingen: Max Niemeyer Verlag.

HOWARTH, P. (1998). The phraseology of learners' academic writing. In A.P. Cowie (ed.). *Phraseology: theory, analysis and applications*. Oxford: Oxford University Press: 161-186.

HYLAND, K. (2002). Specificity revisited: how far should we go now? *English for Specific Purposes*, 21: 385-395.

HYLAND, K. and TSE, P. (2007). Metadiscourse in academic writing: a reappraisal. *Applied Linguistics*, 25(2): 156-177.

LAUFER, B. and LEVITZKY-AVIAD, T. (2006). Examining the effectiveness of 'bilingual dictionary plus' – a dictionary for production in a foreign language. *International Journal of Lexicography*, 19(2): 135-155.

MAN, J.P., WEINKAUF, J.G., TSANG, M. and SIN, D.D. (2004). Why do some countries publish more than others? An international comparison of research funding, English proficiency and publication output in highly ranked general medical journals. *European Journal of Epidemiology*, 19: 811-817.

MOON, R. (2008). Dictionaries and collocation. In S. Granger and F. Meunier (eds). *Phraseology. An Interdisciplinary Perspective*. Amsterdam: Benjamins: 313-336.

MUNGRA, P. and WEBBER, P. (2010). Peer review process in medical research publications. Language and content comments. *English for Specific Purposes*, 29(1): 43-53.

NESSELHAUF N. (2005). *Collocations in a learner corpus*. Amsterdam: Benjamins.

PAQUOT, M. (2008). Exemplification in learner writing: A cross-linguistic perspective. In F. Meunier and S. Granger (eds). *Phraseology in Foreign Language Learning and Teaching*. Amsterdam and Philadelphia: Benjamins: 101-119.

PAQUOT, M. (2010). *Academic Vocabulary in Learner Writing*. London and New York: Continuum.

PAJZS, J. (2009). On the possibility of creating multifunctional lexicographical databases. In H. Bergenholtz, S. Nielsen and S. Tarp (eds). *Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*. Bern: Peter Lang: 327-354.

PECMAN, M. (2008). Compilation, formalisation and presentation of bilingual phraseology: problems and possible solutions. In F. Meunier and S. Granger (eds). *Phraseology in Foreign Language Learning and Teaching*. Amsterdam and Philadelphia: Benjamins: 203-222.

RUNDELL, M. (2007). The dictionary of the future. In S. Granger (ed.). *Optimizing the role of language in technology-enhanced learning.* Proceedings of the expert workshop organized in Louvain-la-Neuve (Belgium), 4-5 October 2007, 49-51. http://juppiter.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/Downloads/GRANGER-SYLVIANE-2007.pdf.

SIEPMANN, D. (2005). *Discourse markers across languages: a contrastive study of second-level discourse markers in native and non-native text with implications for general and pedagogic lexicography*. London and New York: Routledge.

SOBKOWIAK, W. (2002). What can be but is not (and why) in learners' MRDs. *Teaching English with Technology*, 2(3). Available at http://www.iatefl.org.pl/call/j_article9.htm.

TARP, S. (2008). *Lexicography in the Borderland between Knowledge and Non-Knowledge*. Tübingen: Niemeyer. (*Lexicographica Series Maior*, 134).

TARP, S. (2009a). Beyond lexicography: New visions and challenges in the information age. In H. Bergenholtz, S. Nielsen and S. Tarp (eds). *Lexicography at a Crossroads. Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*. Bern: Peter Lang: 17-32.

TARP, S. (2009b). Reflections on data access in lexicographic works. In S. Nielsen and S. Tarp (eds). *Lexicography in the 21st Century*. Amsterdam / Philadelphia: Benjamins: 43-62.

VARANTOLA, K. (2002). Use and usability of dictionaries: common sense and context sensitivity? In M.-H. Corréard (ed.). *Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins*. Euralex: 30-44.

Evaluation of an automatic process for specialized web corpora collection and term extraction for Basque

Antton Gurrutxaga, Igor Leturia, Eli Pociello,
Xabier Saralegi, Iñaki San Vicente[1]
Elhuyar Foundation

Abstract

In this paper we describe the processes for collecting Basque specialized corpora in different domains from the Internet and subsequently extracting terminology out of them, using automatic tools in both cases. We evaluate the results of corpus compiling and term extraction by making use of a specialized dictionary recently updated by experts. We also compare the results of the automatically collected web corpus with those of a traditionally collected corpus, in order to analyze the usefulness of the Internet as a reliable source of information for terminology tasks.

Keywords: web as corpus, automatic term extraction, Basque.

1. Motivation

The traditional process for building corpora – out of printed texts, following some selection criteria, linguistically tagged and indexed, etc. – is a very laborious and costly one, so corpora built in this way are not as large or abundant as we would like them to be, and even less so in specialized domains. So in recent years the web has been used increasingly for linguistic research, both via tools like WebCorp (Kehoe and Renouf 2002) or CorpEus (Leturia *et al.* 2007a) that query search engines directly and show concordances, or via tools that use the Internet as a source of texts for building corpora to be used in the classic way, after linguistic tagging and indexation (Ferraresi *et al.* 2008).

Although the use of the web as a source for building linguistic corpora has its detractors, this approach offers undeniable advantages (Kilgarriff and Grefenstette 2004):

[1] R&D department, Elhuyar Foundation, {a.gurrutxaga,i.leturia,e.pociello,x.saralegi,i.sanvicente}@elhuyar.com

- The corpora that can be obtained are much larger.
- The cost of the automatic building processes is much smaller.
- The web is constantly up to date.

On the other hand, the development of terminological resources is essential for any language that aims to be a communication tool in education, industry, etc. The automation of the term extraction process is a condition for this task to be carried out at a reasonable cost taking large samples of real texts as a data source (Ahmad and Rogers 2001).

If all this is true for any language, it is even more so in the case of a less-resourced language like Basque, so the automation of corpus compilation and terminology extraction processes is very attractive indeed.

2. Corpus collection

The compilation of specialized corpora from the Internet is performed by using an automatic tool (Leturia *et al.* 2008) that gathers the documents via the standard method of search engine queries (Baroni and Bernardini 2004).

The system is fed with a sample mini-corpus of documents that covers as many sub-areas of the domain as possible – 10 to 20 small documents can be enough, depending on the domain. A list of seed terms is automatically extracted from it, which can be manually edited and improved if necessary. Then combinations of these seed words are sent to a search engine, using morphological query expansion and language-filtering words to obtain better results for Basque (Leturia *et al.* 2007b), and the pages returned are downloaded.

Boilerplate is stripped off the downloaded pages (Saralegi and Leturia 2007) which are then passed through various filters:

- Size filtering (Fletcher 2004)
- Paragraph-level language filtering
- Near-duplicate filtering (Broder 2000)
- Containment filtering (Broder 1997)

A final topic-filtering stage is also added, using the initial sample mini-corpus as a reference and using document similarity techniques (Saralegi and Alegria 2007) based on keyword frequencies (Sebastiani 2002). A manual evaluation of this tool showed that it could obtain a topic precision of over 90%.

3. Terminology extraction

Term extraction is carried out using Erauzterm, an automatic terminology extraction tool for Basque (Alegria *et al.* 2004a), which combines both linguistic and statistical methods.

First, a lemmatizer and POS tagger for Basque (Aduriz *et al.* 1996) is applied to the corpus. Then the most usual Noun Phrase structures for Basque terms are detected (Alegria *et al.* 2004b) to obtain a list of term candidates. Term variants are linked to each other by applying some rules at syntagmatic and paradigmatic level. After this normalization step, statistical measures are applied in order to rank the candidates. Multiword terms are ranked according to their degree of association or unithood using Log Likelihood Ratio or LLR (Dunning 1994). Single word terms are ranked according to their termhood or divergence with respect to a general domain corpus, also using LLR. Then those candidates that reach a threshold are chosen. A manual evaluation of the tool reported a precision of 65% for multiword terms and 75% for single word terms for the first 2,000 candidates.

The tool also offers a graphical interface which allows the user, if necessary, to explore, edit and export the extracted terminology.

4. Experiments and evaluation

4.1. Experiments

We used the tools and systems described above to collect three specialized corpora and to obtain term lists from them, and we evaluated the results.

The domains chosen were Computer Science, Biotechnology and Atomic & Particle Physics. The collection of the corpora from the Internet did not have a target size, because the Internet in Basque is not as big as that in other languages, and the number of pages we would want to collect for a particular domain might not exist. So we simply launched the collecting processes and stopped them when the growing speed of the corpora fell to almost zero, thus obtaining corpora that were as large as possible. Then we applied the terminology extraction process to the corpora and obtained three term candidate lists. These lists were automatically validated against a recently compiled specialized dictionary, *Basic Dictionary of Science and Technology* (http://zthiztegia.elhuyar.org), which contains 25,000 terms. The best ranked ones of the remaining candidates were manually evaluated by experts to decide if they were terms or not.

Table 1 shows the size of the corpora obtained, the number of terms extracted and the number of terms validated manually or by the dictionary, for each of the three domains.

Corpus	Atomic and Particle Physics	Computer Science	Biotechnology
Sample corpus size	32 docs, 26,164 words	33 docs, 34,266 words	55 docs, 41,496 words
Obtained corpus size	320,212	2,514,290	578,866
Extracted term list size	46,972	163,698	34,910
Dictionary validated	6,432	8,137	6,524
Manually evaluated	1,147	905	628
Terms	887	513	432
Not terms	260	392	196

Table 1. Corpus and term list sizes obtained for each of the three domains

4.2. Evaluation

We evaluated the domain precision of the lists obtained from the Internet, by analyzing the distribution of the terms across the domains, taking the domains of the specialized dictionary as a reference. The results of this evaluation are shown in Figure 1, where we can observe that all three lists show peaks in or around their respective domains, which proves that the corpora are indeed specialized and that the term lists automatically extracted belong mainly to the desired domains.

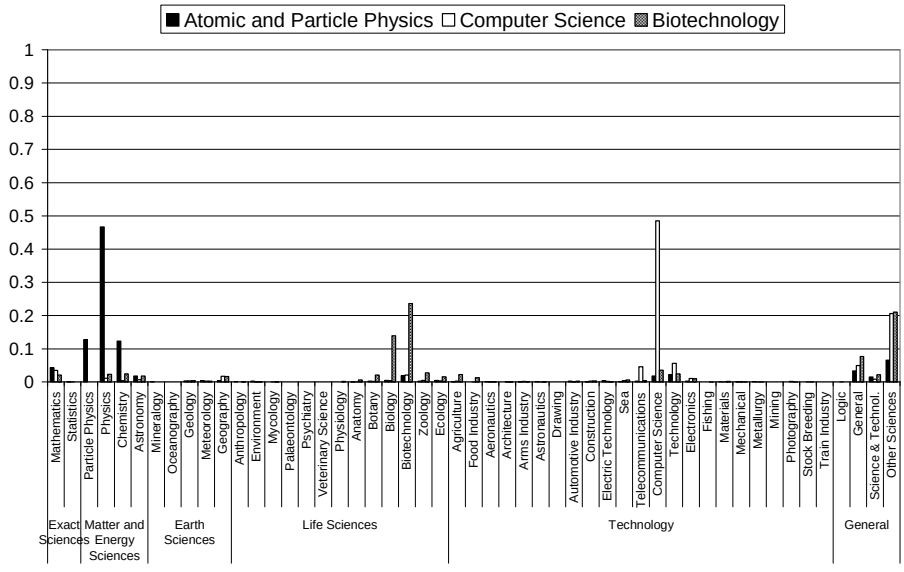


Figure 1. Domain distribution of the extracted term lists

The precision of the extracted term lists, that is, the percentage of the extracted terms that really belonged to the desired domain, was also evaluated. Figure 2 shows the evolution of this precision as the number of candidate terms grows. Here we can observe that the results are different for each of the domains. As a general rule, we can say that pure sciences perform better than technologies, which might indicate that these domains are more "terminologically dense", although we cannot be sure about this, because it could also be due to the different nature – extension, diversity, production – of the domains. Besides, we believe that the seed document selection might also affect the quality of the resulting corpora and term lists.

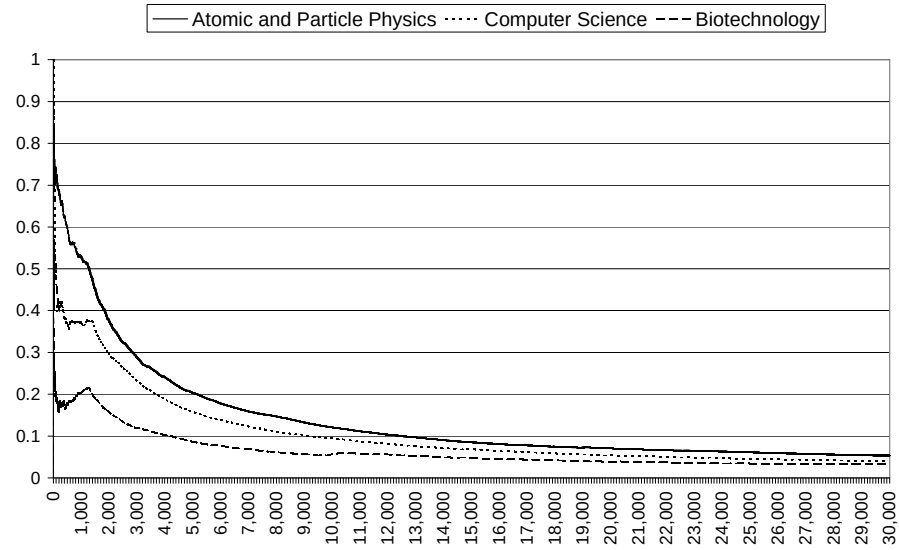


Figure 2. Domain precision of the extracted term lists

Also, the size of the collected corpora does not seem so important as far as the term extraction task is concerned: the Atomic and Particle Physics corpus achieves better results than the Biotechnology one, the former being almost half the size of the latter (Table 1). As we have already pointed out, the nature of the domain is more important.

We also compared the extracted term lists with the lists on the domains of a specialized dictionary compiled and recently updated by experts, and look at the recall, that is, the percentage of the dictionary achieved, and the number of new terms extracted that were not in the dictionary. These two pieces of data are shown in Figures 3 and 4. By looking at the recall, we could draw the conclusion that the corpus building process is not good enough for compiling a quality dictionary, but we will see later that a traditional corpus does not do better. The use of corpora lacking representativeness could be put forward as a reason for that flaw. But another possible explanation for this fact could lie in the current situation of Basque terminology and text production. Although Basque began to be used in Science and Technology thirty

years ago, it cannot be denied that there is a given amount of highly specialized terminology that is published *ex novo* in dictionaries, with little document support if any. That could be the reason why several terms chosen by experts and published in the dictionary do not occur in either of the two corpora. However, we can see in Figure 5 that many new terms appear, so the process proposed is definitely interesting for enriching or updating already existing specialized dictionaries.

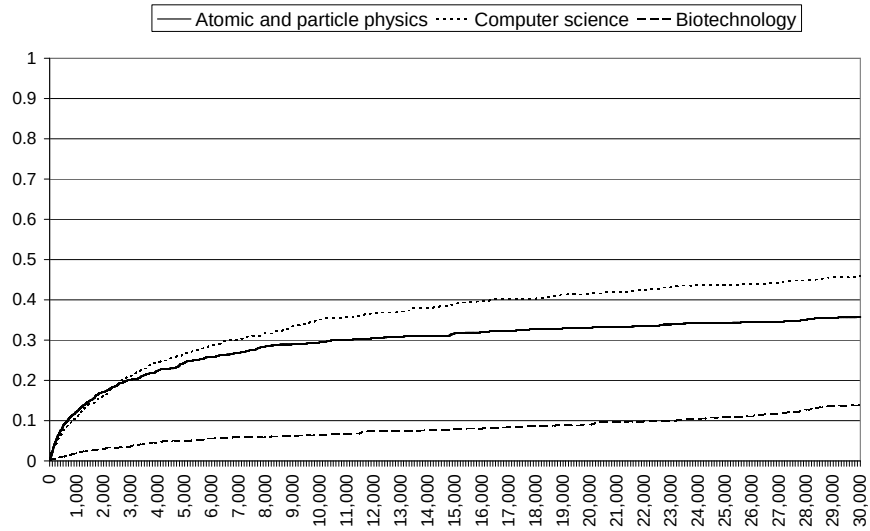


Figure 3. Recall of the extracted term lists compared with the dictionary

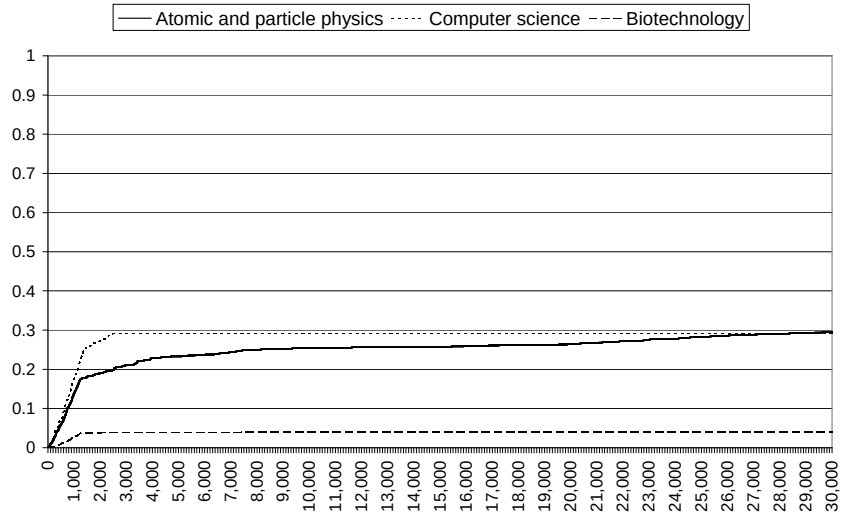


Figure 4. New terms in the extracted term lists that were not in the dictionary

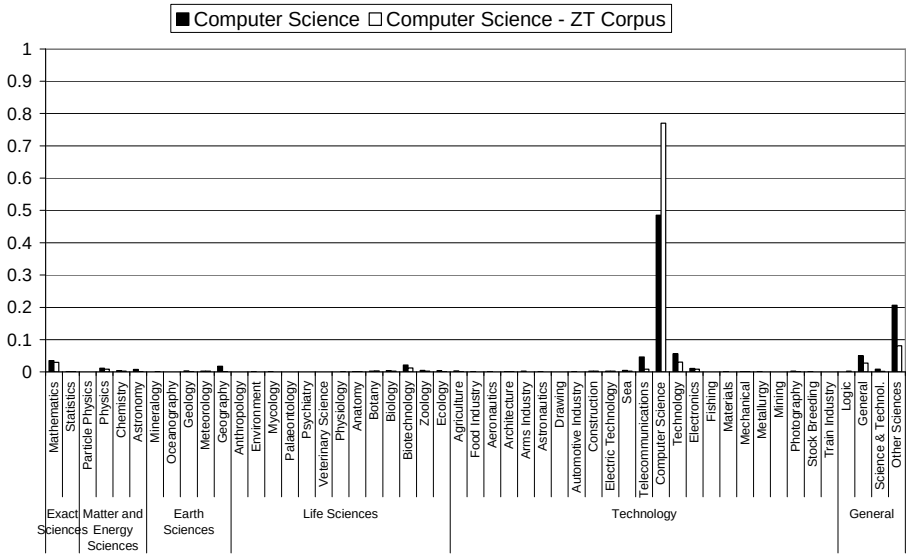


Figure 5. Domain distribution of the extracted term lists

Finally, we compared the term list extracted from a corpus automatically collected from the web with the term list extracted from a classical corpus. So a sub-corpus of the Computer Science domain was extracted from a traditional corpus, the ZT Corpus (Areta *et al.* 2007; http://www.ztcorpusa.net), and terminology was extracted with the same method used with the Computer Science web corpus. Then both lists were compared. Table 2 shows data on these two corpora and their respective term lists.

Corpus	**Computer Science**	**Computer Science – ZT Corpus**
Sample corpus size	33 docs, 34,266 words	-
Obtained corpus size	2,514,290	332,745
Extracted term list size	163,698	24,283
Dictionary validated	8,137	3,389
Manually evaluated	905	1,022
Positive	513	479
Negative	392	543

Table 2. Corpus and term list sizes obtained for the web and traditional corpora

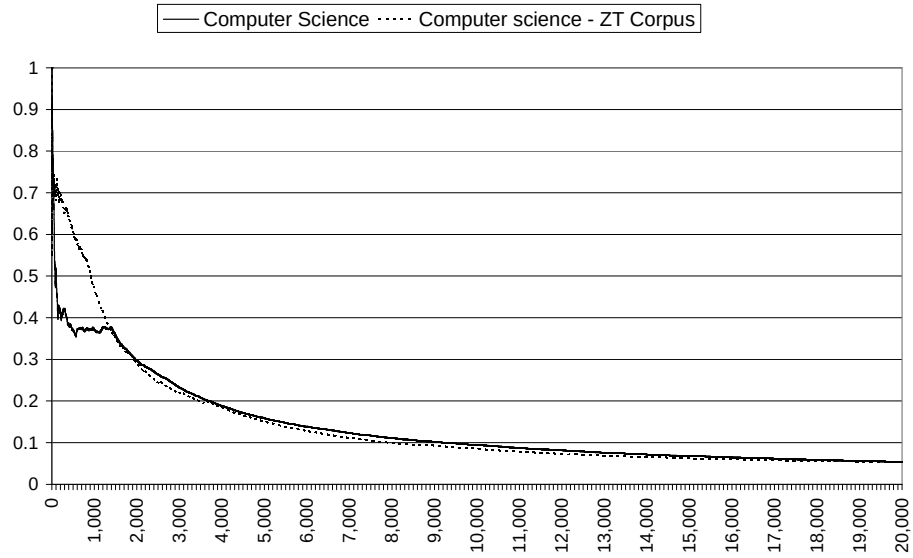


Figure 6. Domain precision of the extracted term lists

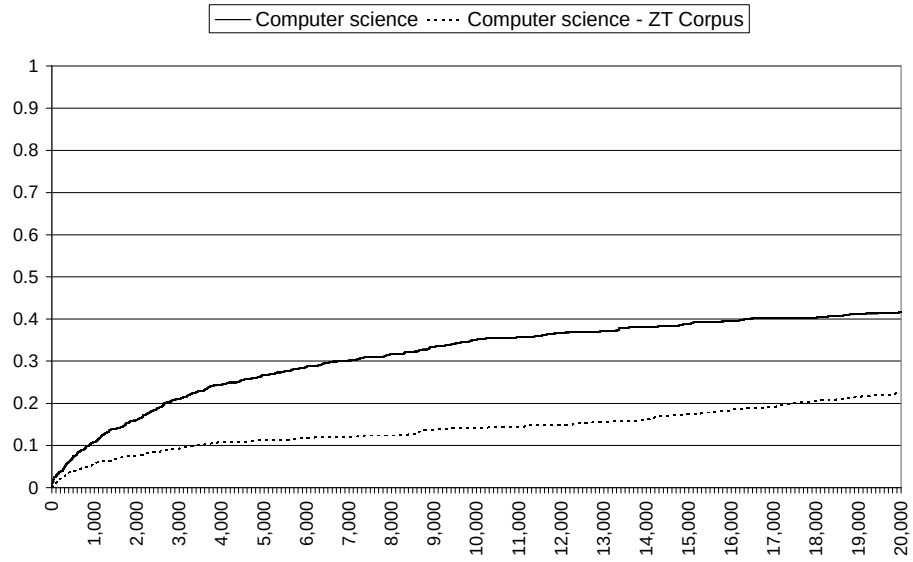


Figure 7. Recall of the extracted term lists compared with the dictionary

Figures 5, 6, 7 and 8 show, respectively, the domain distribution, domain precision, recall compared with the dictionary and new terms that were not in the dictionary of

the two extracted term lists. They prove that we can obtain similar or, in some aspects, even better results with the automatic corpus collection process. As the cost is much lower, we believe that the process proposed in the paper is valid and very interesting for terminological tasks.

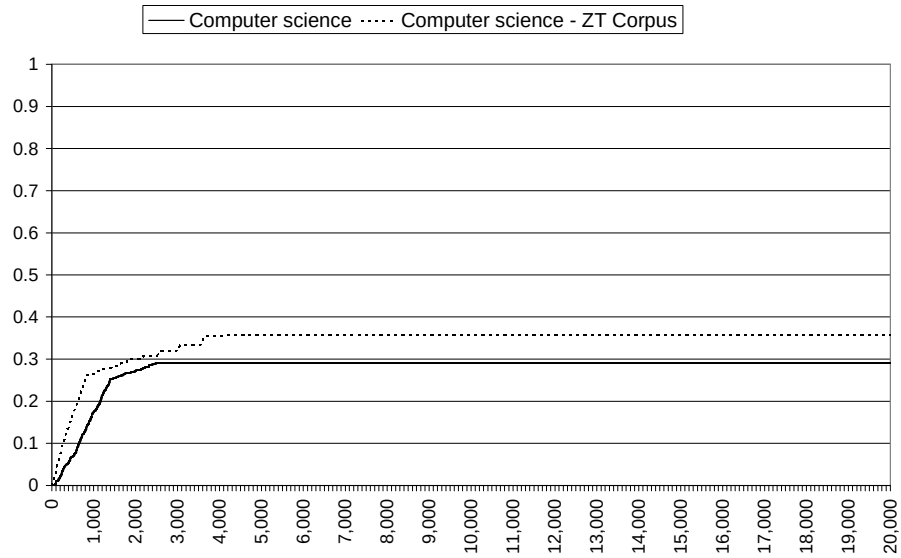


Figure 8. New terms in the extracted term lists that were not in the dictionary

5. Conclusions

The distribution and domain precision graphs prove that the corpora and term lists obtained are indeed specialized on the desired domains and lead us to believe that the automatic corpus collection and term extraction process can be valid for terminology tasks.

The evaluation also shows that results almost as good as from a traditional corpus can be obtained regarding precision or new terms, and even better in the case of recall.

Overall, the evaluation results are encouraging and indicate that acceptable results can be obtained with much less work than by means of a completely manual process.

References

ADURIZ, I., ALDEZABAL, I., ALEGRIA, I., ARTOLA, X., EZEIZA, N. and URIZAR, R. (1996). EUSLEM: A Lemmatiser / Tagger for Basque. In *Proceedings of EURALEX'96*. Göteborg: EURALEX: 17-26.

AHMAD, K. and ROGERS, M. (2001). Corpus Linguistics and Terminology Extraction. In S.E. Wright and G. Budin (eds.) *Handbook of Terminology Management*, vol. 2. Amsterdam: John Benjamins: 725-760.

ALEGRIA, I., GURRUTXAGA, A., LIZASO, P., SARALEGI, X., UGARTETXEA, S. and URIZAR, R. (2004a). An Xml-Based Term Extraction Tool for Basque. In *Proceedings of the 4th International Conference on Language Resources and Evaluations (LREC)*. Lisbon: ELRA: 1733-1736.

ALEGRIA, I., GURRUTXAGA, A., LIZASO, P., SARALEGI, X., UGARTETXEA, S. and URIZAR, R. (2004b). Linguistic and Statistical Approaches to Basque Term Extraction. In *Proceedings of GLAT 2004: The production of specialized texts*. Barcelona: ENST Bretagne: 235-246.

ARETA N., GURRUTXAGA, A., LETURIA, I., ALEGRIA, I., ARTOLA, X., DÍAZ DE ILARRAZA, A., EZEIZA, N. and SOLOGAISTOA, A. (2007). ZT Corpus: Annotation and tools for Basque corpora. In *Proceedings of Corpus Linguistics*. Birmingham: University of Birmingham.

BARONI, M. and BERNARDINI, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of the 4th International Conference on Language Resources and Evaluations (LREC)*. Lisbon: ELRA: 1313-1316.

BRODER, A.Z. (1997). On the resemblance and containment of documents. In *Proceedings of Compression and Complexity of Sequences 1997*. Los Alamitos: IEEE Computer Society: 21-29.

BRODER, A.Z. (2000). Identifying and filtering near-duplicate documents. In *Proceedings of Combinatorial Pattern Matching: 11th Annual Symposium*. Montreal: Springer: 1-10.

DUNNING, T. (1994). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1): 61-74.

FERRARESI, A., ZANCHETTA, E., BARONI, M. and BERNARDINI, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus workshop*. Marrakech: ELRA: 47-54.

FLETCHER, W.H. (2004). Making the web more useful as a source for linguistic corpora. In U. Connor and T. Upton (eds). *Corpus Linguistics in North America 2002*. Amsterdam: Rodopi: 191-205.

KEHOE, A. and RENOUF, A. (2002). WebCorp: Applying the Web to Linguistics and Linguistics to the Web. In *Proceedings of the WWW2002 Conference*. Honolulu: W3C.

KILGARRIFF, A. and GREFENSTETTE, G. (2004). Introduction to the special issue on the Web as corpus. *Computational Linguistics*, 29: 333-348.

LETURIA, I., GURRUTXAGA, A., ALEGRIA, I. and EZEIZA, A. (2007a). CorpEus, a web as corpus tool designed for the agglutinative nature of Basque. In *Proceedings of the 3rd Web as Corpus workshop*. Louvain-la-Neuve: Presses Universitaires de Louvain: 69-81.

LETURIA, I., GURRUTXAGA, A., ARETA, A., ALEGRIA, I. and EZEIZA, A. (2007b). EusBila, a search service designed for the agglutinative nature of Basque. In *Proceedings of Improving non-English web searching (iNEWS'07) workshop*. Amsterdam: SIGIR: 47-54.

LETURIA, I., SAN VICENTE, I,. SARALEGI, X. and LOPEZ DE LACALLE., M. (2008). Basque specialized corpora from the web: language-specific performance tweaks and improving topic precision. In *Proceedings of the 4th Web as Corpus workshop*. Marrakech: ELRA: 40-46.

SARALEGI, X. and ALEGRIA, I. (2007). Similitud entre documentos multilingües de carácter científico-técnico en un entorno web. *Procesamiento del Lenguaje Natural*, 39: 71-78.

SARALEGI, X. and LETURIA, I. (2007). Kimatu, a tool for cleaning non-content text parts from HTML docs. In *Proceedings of the 3rd Web as Corpus workshop*. Louvain-la-Neuve: Presses Universitaires de Louvain: 163-167 (*Cahiers du Cental*, 4).

SEBASTIANI, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1): 1-47.

Elliptical arguments: a problem in relating meaning to use

Patrick Hanks[1]
Charles University in Prague

Abstract

Corpus lexicographers working in the tradition of John Sinclair (of whom the present author is one) argue that electronic dictionaries of the future will have a duty to pay close attention to phraseology and to phraseological meaning in text. To do this, they will need to make a distinction between normal patterns of use of words and exploitations of normal uses, such as freshly coined metaphors, used for rhetorical and other effects. Electronic dictionaries will report the norms of phraseology associated with phraseological patterns instead of or as well as word meaning in isolation. The foundations for this approach to lexical analysis are explored in the Theory of Norms and Exploitations (Hanks, in press). The present paper starts by discussing the relationship between valency and collocation and goes on to discuss a particular problem in collocational and valency analysis, namely the effect on clause meaning of omitted arguments. For example, the verb 'fire', with a human subject, has two main meanings: "to discharge a projectile from a firearm" and "to dismiss a person from employment". However, if the direct object is omitted, only the first sense of the verb can be activated, not the second. The paper investigates some corpus-based examples of elliptical arguments and discusses their semantic implications.

Keywords: phraseology, patterns, valency, collocations, semantic types, norms, exploitations, ellipsis.

1. Introduction: dictionaries and phraseology

Electronic dictionaries of the future will be much in demand – for computational, pedagogical, and other applications – if they can be used as resources for mapping word meaning systematically onto word use. Research in computational linguistics has shown that algorithms using electronic versions of current dictionaries designed for human users cannot achieve this goal (see, for example, Ide and Wilks 2006, Hanks 2008). Nor are results using hierarchical ontologies such as WordNet any better (see Nirenburg 2007). Such resources are plausible enough for human users, but they fail to meet the challenges of mapping meaning systematically onto words in use in free text for computation. One reason for this is that research questions have often been formulated on false assumptions. Fillmore (1975) warns against 'checklist' theories of meaning. The notion that a word represents a list of senses that can be disambiguated

[1] Institute of Formal and Applied Linguistics (UFAL), Charles University in Prague, patrick.w.hanks@gmail.com

by some procedure or other is a dangerously crude generalization, encouraged by superficial inspection of traditional dictionaries and WordNet synsets. Fillmore was one of the very first linguists to draw attention to an alternative theory, namely that meanings, like certain other linguistic phenomena, are best interpreted probabilistically, in terms of best match to a prototype, rather than by satisfaction of sets of necessary and sufficient conditions.

Meanings are associated with patterns (Hunston and Francis 2000), constructions (Goldberg 1995, 2006), or 'phrasemes' (Mel'čuk 1988), as well as with words. Corpus analysis, using tools such as the Sketch Engine (Kilgarriff *et al.* 2004), shows that usage is both variable and highly patterned. The collocational preferences of lexical items within a phraseme must be analysed statistically (Sinclair 1966, 1987, 1991; Church and Hanks 1990; Kilgarriff *et al.* 2004), if the meaning potential of the lexical item is to be understood. In the theory of norms and exploitations, 'TNE' (Hanks 1994 and in press), I propose that some variations are themselves norms, known as 'alternations'. Others are rule-governed exploitations of norms.

In this paper, I explore the idea that a meaning may be associated, not only with particular patterns and pattern elements (valencies and collocations), but also, in certain circumstances, with the absence of a particular pattern element.

2. What is a norm? What is a pattern?

Normally, each content word used by a speaker of a language to make a meaning is used in conformity to one or other of the patterned norms in the language for using that word. However, occasionally speakers and writers exploit norms.

A norm, in the sense denoted here, is a use of a word that conforms to any one of several semantically motivated syntagmatic patterns with which that word is associated – or, to borrow Michael Hoey's (2005) term, 'primed'. These syntagmatic patterns can be discovered by corpus analysis, but not by introspection or the invention of evidence. Invented evidence distorts.[2]

For verbs, a pattern consists of the verb itself together with its arguments, the latter being populated by lexical sets. A lexical set is a group of semantically related content words – synonyms, antonyms, hyponyms, etc. – that are found in the same valency relation to a collocate – the verb – in a sentence: the words of a lexical set activate the same meaning of the collocate. The words in a lexical set share some semantic feature, in most cases, this is their semantic type.

A lexical set may consist of just a single word: for example, the meaning of the idiomatic expression ***grasp the nettle***, 'to deal resolutely with a difficult issue',

[2] The role of intuitions is a subject that has excited much controversy in both lexicography and linguistics. The view taken here is that of course intuitions are absolutely necessary for interpreting data, but that it is neither necessary nor desirable to invent examples. An exception to this stricture is the invention of contrastive examples for explanatory purposes, as in examples 9a-9c below.

depends for successful realization on the presence of the particular lexical item ***nettle*** in colligation with a verb of seizing. No other word will do: you cannot, for example, realize the same meaning by talking about 'grasping the poison ivy' or 'grasping the scorpion'. At the other extreme, a lexical set may be huge and indeed open-ended. Thus, the semantic type [[Human]] comprises an immense and indeed in principle infinite number of lexical items. This is a lexical set that is found as the subject of most sentences and very often in the object and adverbial slots, too – for much if not most human discourse is about who did what to whom. As a general rule, the smaller the lexical set, the more precise its effect on the meaning of the pattern. Again, it must be emphasized that lexical sets are part of the system of the language, to be discovered by empirical corpus analysis. A person cannot just decide to invent a new lexical set at will.

Some lexical sets alternate in given contexts with other, more salient lexical sets: for example, in the subject slot for verbs denoting cognitive procedures, the semantic type [[Human Institution]] regularly alternates with the prototypical type [[Human]]. In actual fact, it is humans who think and say things, but it is also very normal to find sentences in which a government or other human institution is *said* to say or think something.

3. Exploitations of norms

A few utterances may be described as exploitations of norms. These are the truly creative utterances in a language. There is no firm boundary between the category of exploitations and the category of alternations. The two categories shade imperceptibly into each other. A sentence that, for the utterer, is merely an alternation or a domain-specific norm may strike some hearers or readers as a particularly creative piece of phraseology.

In doing corpus analysis of a word, most corpus lines can be classified as realizations of a particular pattern; others are alternations; a few are exploitations. As a rule of thumb, a corpus analysis in which more than about 10% of the uses of the word being analysed are classed as exploitations is worth re-examining to see whether some secondary norm or alternation has been missed.

An example may help to clarify what has just been said. The English verb ***hazard*** has two patterns:

A. [[Human]] hazard {guess}

B. [[Human]] hazard [[Entity = Valued]]

Sentence 1 below illustrates the most normal use of the verb ***hazard*** (in the more common of the two patterns); sentence 2 illustrates a simple alternation of the norm, in which other terms denoting speech acts or propositions (description, definition) are used in place of the prototypical direct object, ***guess***, and sentence 3 illustrates an exploitation, which will be discussed in more detail towards the end of this paper.

1. I can only hazard a guess at what it must have been like to sail in a typical convoy.[3]
2. Could you hazard a description or definition of the id?
3. I hazarded various Stuartesque destinations like Florida, Bali, Crete and Western Turkey.

All this implies that use of the words of a natural language is rule-governed, but not by a single rule system; rather, there is a 'double helix' of two interacting rule systems: rules for using words normally, and rules for exploiting norms. It is necessary to move away from 'Lego-set' theories of meaning in language, in which words are thought to be put together like children's toy bricks in order to make meaningful propositions. Such theories are simplistic and lack descriptive adequacy. If we adopt a more sophisticated approach, in which meanings are associated with patterns or constructions, as well as with words in isolation, the semantic interaction between conventional phraseological norms and creative exploitations of those norms will be seen to play a crucial role.

A problem for linguists attempting to break away from Lego-set theories of language is: what to replace it by? It is now clear from over two decades of work in corpus linguistics that natural languages are probabilistic systems in which patterns are clearly present but highly variable. In order to describe a pattern, some degree of idealization is inevitable and difficult choices must be made. For example, should the direct object of ***hazard*** be represented as the single lexical item {guess}, as in A above? This particular lexical item is highly prototypical, being found as the direct object in more than 50% of all uses of this verb. Alternatively, should it be represented more generally, as the semantic type [[Speech Act]]? The latter would admit not only guesses, conjectures, definitions, and descriptions, but also all sorts of other speech acts, some of which are highly unlikely: for example a command is a speech act, but you do not normally *hazard* a command. The traditional lexicographical solution, 'something', is not actually wrong, but it is of course severely underspecified.

4. Valency and semantic types

If meanings are to be associated with patterns, rather than merely with words in isolation, then detailed corpus analysis is needed to discover the patterns with which each word is associated and the strength of the association.

Each content word of a language is used in one or more valency structures, as has been described for English in considerable detail by Herbst *et al.* (2005). In many cases, different valencies are associated with different senses. For example, consider the verb ***shower.*** If you ***shower*** (intransitive), the meaning is that you wash your body – and,

[3] In this paper, examples are taken from the British National Corpus (BNC), unless otherwise stated. Authentic uses of words (taken from corpora and other texts) are printed in roman, while invented examples (used mainly for contrastive purposes) are in italics.

usually, your hair – by standing under a device that emits a spray of water, thus getting wet all over. On the other hand, if you ***shower*** someone with something, then the default assumption is of a completely different type of action, and there is no implication that anyone gets wet.

If the verb ***shower*** is intransitive, *water* may assumed by default – it is quite unnecessary to state that you shower *with water*. *Showering with water* is vanishingly rare phraseologically. There are no examples of this expression in the BNC, and I was able to find only a couple of examples by surfing Google, for example, 4 – a caption to a photograph, where the point is that the water is not doing what water in a shower normally does, *i.e.* coming out of a showerhead in a bathroom.

4. Carlos Rodriguez, 28, enjoys a shower with water coming from a public stream.

What about transitive uses of this verb? Normally, as soon as you introduce a direct object – *showering someone* – you commit yourself to an adverbial as well: you shower someone *with something*, such as gifts or abuse. A transitive use without an adverbial, for example *Janet showered her children*, is possible in English, and it is perfectly grammatical but it is not normal. In fact it is exceedingly rare – so rare that I have been reduced to inventing an example, because I could not find one in the corpora I am studying. The implicature, of course, is that Janet washed the children, not that she gave them a lot of gifts, praise, or abuse.

To give a fully adequate account of the semantically distinct patterns in which ***shower***, as a verb, participates, together with their meanings, it is necessary to go beyond the valency structure and specify the semantic type of at least some of the arguments.

5. Boris showered the woman with presents.
6. Lauren Bacall, Bianca Jagger, Claire Bloom, Linda Thorson and Lionel Blair were among the stars who showered him with praise.
7. 300 yobs showered police with broken bottles and bricks.
8. Mount Pinatubo erupted …, showering Manila … with quantities of ash and grit.

The verb ***shower*** has identical valency structures in 5-8, and in 5, 6, and 7 the semantic type of both subject and object is [[Human]]. However, in 5, the semantic type of the prepositional object is [[Gift]], with the result that the verb must be interpreted as a Giving event, whereas in 6 it is a Speech-Act event and in 7 it is a Throwing event. Finally, if the semantic type of the subject is [[Volcano]] or [[Explosion]], as in 8, rather than [[Human]], the event type of ***shower*** is likewise Throwing.

Many verbs are like ***shower***: the meaning is strongly influenced by the collocates. Other verbs, like ***organize***, are more terminological. The meaning of this verb is something like 'to put into good order', and the event type denoted is much less dependent on the collocates than in the case of verbs like ***shower***. Typically, a [[Human]] organizes almost anything, ranging from an [[Event]] such as a birthday party to a mass of [[Physical Object]]s such as a pile of papers on a desk top. And then

there are machines and procedures that organize events or groups of entities, still with much the same meaning – *i.e.* much the same event type.

5. Focusing on ellipsis

We can now turn to one particular problem in corpus pattern analysis, namely ellipsis or omission – *i.e.* patterns and exploitations in which an expected argument is not explicitly realized. Consider the verb ***fire***. Over a dozen different patterns of normal use of this verb can be distinguished. They are nearly all transitive, but one of them has an intransitive alternation. Some of these patterns activate similar meanings; others activate quite different meanings. In the most basic pattern, illustrated here by sentence 9, the meaning is 'cause a firearm to discharge a projectile'. This contrasts with other patterns of the same verb, activating other meanings, namely in 10 'to stimulate or excite', in 11 'to expose to heat in a kiln', and in 12 'to dismiss from employment'.

9. I was in a place once when a man fired a gun at me and I did not like it at all.
10. Active citizenship has already fired the imagination of many people.
11. Fashioning and firing a pot does not affect the clay composition.
12. General Avril fired four lieutenant-colonels.

In these examples, the semantic types of the arguments activate different senses of the verb. In 9 the direct object is [[Firearm]], in 10 it is [[Mental Activity]], in 11 it is [[Ceramic]] and in 12 it is another [[Human]]. Each of these direct objects correlates with certain dependencies and the semantic types of other arguments; for example, in 9 and 12 there is a correlation with the subject, [[Human]], but 9, unlike 12, also correlates with an adverbial of direction – 'at me'. The direct object in 10 typically governs a dependent possessive – here, 'of many people' – and correlates with a subject of semantic type [[Abstract Entity]].

It would be very convenient if natural language always behaved in the way suggested by these carefully selected contrastive examples. However, it does not. In ordinary language use, there are some circumstances in which an argument can be omitted, while in other cases it cannot. These omissions rarely bother human readers and hearers, because the speaker or writer correctly judges the omitted item to be 'obvious'. Only obvious arguments can be elided. The elided argument is taken to be common knowledge and therefore does not need to be stated. Electronic dictionaries of the future, however, must account for the circumstances under which ellipsis (optional omission) is possible.

With 9, both the direct object and the adverbial of direction are optional. One can say:

9a. *I was in a place once when a man fired at me and I did not like it at all*

or:

9b. *I was in a place once when a man fired a gun and I did not like it at all.*

In 9a the [[Firearm]] is omitted and in 9b the [[Target]] is omitted. In an appropriate context, one can even omit both arguments, saying:

9c. *He fired.*

Here, the meaning must be that he fired a gun. 9c is quite unambiguous, even though the verb is polysemous and there appears to be no disambiguating context. If a verb is polysemous, ellipsis is an alternation found only with one or more of the most literal pattern(s). Any computational linguistic procedure seeking to assign a meaning or a translation to ***fire*** by analysis of context must look for a direct object and, not finding one, can conclude that there is a very high degree of probability that the meaning is 'discharge a projectile from a firearm towards a target'.

True to the Gricean maxim of quantity, speakers and writers generally do not say more than is necessary. Omission of words – ellipsis – is a very common phenomenon in ordinary language use. This can lead to violations of strict principles of syntactic well-formedness, although it is consistent with the principle of textual well-formedness (Sinclair 1984). As a result, a sentence taken in isolation from the context in which it is embedded may seem to be very ambiguous. Consider 13, a sentence that has been artificially isolated by being taken out of context.

13. Later that morning he changed.

The interpretation of ***changed*** in this sentence is dramatically affected, not by the complementation, but by the wider context. To see this, imagine that 13 has preceding context as in 13a, then imagine 13b, and then imagine 13c.

13a. *At breakfast he was still wearing a black tie and crumpled dinner jacket from the night before.* Later that morning he changed.

13b. *At breakfast he greeted us with a cheerful grin and seemed not to have a care in the world.* Later that morning he changed.

13c. *He got on at Köln thinking that it was a through train to Berlin, but the ticket inspector told him that it would terminate at Hannover*. Later that morning he changed.

The meaning of ***change*** is completely different in each of these three cases. Whatever the interpretation, which depends on the context established in the text leading up to the clause containing the verb *changed*, sentence 13 exemplifies the very common '**null-object alternation'**, also called the **'object-drop alternation'** or '**unexpressed object alternation**'. A writer can reasonably expect that a reader will proceed sequentially through a text and therefore that the reader can predict what the expected direct object is, which in turns means that the writer does not need to state it explicitly.

Other examples of object ellipsis from BNC are 14-16.

14. This suggests that many small farmers, unable to cultivate successfully, turned to the sale or renting of land. – (BNC) Tessa Cubitt, 1988. *Latin American Society*.

Cultivate is normally a transitive verb, but in 13 the direct object is left unstated, presumably because the writer considers it obvious that what farmers cultivate is the land.

A similar example is 15, from a description of the effect that W. P. Nicholson, a Northern Irish fundamentalist Protestant preacher, had in the University of Oxford when he was invited there as a missionary in the 1920s.

15. In Holy Trinity Church Nicholson abounded in anecdotes, vulgarity, rudeness, emotional appeals, a dogmatism so dogmatic as to frighten. More and more people went to hear this phenomenon in a university of the crudest fundamentalism, which horrified some of the dons as a caricature of Christianity. People who could not bear it walked out.

(BNC) Owen Chadwick, 1991. *Michael Ramsey: a Life*

Frighten is normally a transitive verb: it requires a direct object. When the direct object is omitted, the reader or hearer is left to 'understand' a default direct object, namely anything with the semantic type [[Animate]], but in this context restricted to a subset of animates – human beings who happened to be Christians in Oxford in 1925 and who heard Nicholson's sermons.

In 16, there are two elliptical alternations in a single sentence.

16. We punish too much – and in particular, we imprison too much.

(BNC) J. Dignan, 1992. *The Penal System*

Punish and ***imprison*** are normally transitive verbs, taking both a subject and a direct object with the semantic type [[Human]]. The usual focus is on the person being punished. There is generally also a prepositional *for*-phrase saying what he or she had done that was punishment-worthy – and if it is not actually present, it is certainly implicit. Sometimes, there is also a *to*-phrase saying what penalty or retribution was meted out. But in 16, there is no direct object, no prepositional phrase saying what anyone is punished for, and no mention of a penalty. This alternation, with the absence of the expected direct object and adverbial, has the effect of generalizing the sense of the verb. In this context, who is being punished and for what is deliberately left unstated, and the focus instead is on the general act of punishing.

Another example is the verb ***decline*** in 17. What did the Englishman decline? The sentence does not tell us explicitly, but we can be sure that the answer is somewhere in the preceding context. In 17, it is the antecedent of the pronoun *one* – a cigarette, as it happens.

17. He offered one to Estabrook, who declined.

Sinclair (1991) comments on this verb:

> Whatever is reported as having been declined has already been named, mentioned, or indicated with sufficient clarity; so that the reader, arriving at the word ***declined***, need be in no doubt about what would be a suitable object or infinitive clause.

6. Ellipsis of adverbials

It is not only the direct objects of verbs that can undergo omission as an elliptical alternation.

In the case of many verbs that take a completive-intensive particle, *e.g.* ***calm (down)***, the particle is optional. This type of alternation is found with many verbs denoting processes. Omission of the particle can be regarded as an elliptical alternation; alternatively, its inclusion may be regarded as a pleonastic alternation.

Another kind of ellipsis involves dropping an adverbial under certain conditions, which seem to be verb-specific. This is not to be confused with cases where an adverbial adjunct is entirely optional.

The adverbial valency of verbs in English is the subject of much confusion. This is not surprising, because the facts of the language themselves in particular are confused and confusing. Some adverbials are obligatory; others are optional; and to make matters worse, some obligatory adverbials can be elided! The confusion is made even worse by differences of terminology in competing grammatical traditions. Here, I will attempt to describe the salient facts briefly, with examples – but only insofar as is necessary for effective corpus analysis of the lexicon – using terminology taken eclectically from at least three major traditions. I shall not attempt a full summary of the role of adverbials in these traditions.

First, let us look at a case where an adverbial argument is obligatory. The verb ***put*** is such a verb. Consider 18 and 19.

18. He put the painting on the floor.
19. Put the light here.

Put is one of many verbs in English which, for grammatical and semantic well-formedness, require a valency of three clause roles around it – the person doing the putting, the thing that is put, and the place in which it is put. Standard American dictionaries, which subcategorize ***put*** merely as a transitive verb, with no mention of the adverbial, fail to tell the full story. Using such a dictionary for NLP or language learning must be like trying to run with only one leg, for with this verb ellipsis of the adverbial is impossible. You cannot say, **Put the light* or **Put the painting*. Such an utterance would be both syntactically and textually ill-formed in all imaginable circumstances, lying well beyond the grey area of permissible alternations and exploitations.

Now consider, by way of contrast, the verb ***abstain***. Here, there is considerable variation as regards the presence or absence of an adverbial argument. It is an intransitive verb which, in its canonical form, takes a ***from*** phrase as an adverbial argument, in which the governed noun phrase denotes an [[Activity]], as in 20.

20. I will abstain from discussing these aspects here.

If, as in the case of the first two direct objects of ***abstain*** in 21, the adverbial argument contains a governed noun phrase denoting a [[Physical Object]], the [[Physical Object]] is coerced to having the value of an [[Activity]] most typically.

21. I have kept myself fit all my life, avoiding infections, abstaining from drink, tobacco and hazardous pursuits.

Drink here means 'ingesting alcoholic beverage' and *tobacco* means 'smoking'. The mechanisms of such semantic coercions are described in more detail in Pustejovsky (1995).

The most common use of this verb is not, however, to do with discussions, drinking, smoking, or hazardous pursuits, but rather in political contexts with reference to a vote – if you are entitled to vote, you can abstain from voting. This sense is so common that a second normal pattern of use has developed without an adverbial. If there is no adverbial, the default meaning is 'to deliberately not vote', as in 22.

22. The National People's Congress voted 1,767 to 177 in favour of building the dam, but 664 abstained.

Thus, there are two patterns of normal usage for this verb, associated with different meanings:

A. [[Human]] abstain {from [[Activity]]}

= [[Human]] deliberately does not do [[Activity]]

B. [[Human]] abstain [NO OBJ]

= [[Human]] deliberately does not vote

Unfortunately for linguistic analysis, however, absence of an adverbial does not *necessarily* activate sense B. This is the default meaning if there is no adverbial, but pattern A can also participate in a null-adverbial alternation, as in 23, where both the default implicature and the wider context – not quoted here – make it clear that the speaker is talking about abstaining from drinking alcohol, not abstaining from a vote.

23. The longest period I've **abstained** was two-and-a-half months.

Many other verbs commonly govern prepositional phrases, but these are optional, not obligatory. Typical is ***die***, which is often cited in the linguistics literature as an example of a verb that has only one argument – the subject or 'external argument'.[4] This is correct, even though ***die*** almost always governs one or more prepositional phrases – as in 24, where there are three of them, expressing cause, date, and location. The point is that even though ***die*** rarely occurs without one or more adverbials, all of the adverbials are structurally optional in respect of the meaning of the verb.

24. Bob Fitzsimmons **died** of pneumonia on 22 October 1917, in Chicago, Illinois.

[4] "external" because it is not governed by the verb.

With this verb we encounter a minor theoretical paradox. As a general rule, the whole point of using the verb ***die*** is to mention the date on which someone died, the place where they died, the cause of death, and/or the social, physical, or financial circumstances affecting them when they died – not merely to state that the event took place. All such information is expressed in adverbials. But, as already noted, even though one or more adverbials are almost always found with this verb, it is not obligatory to have one. '*He died*' is not an ill-formed sentence of English in the same way as '*He put*', and it is not a case of contextually licensed ellipsis, like ***abstain*** in 23, because there is no implicit understanding about what the missing adverbial might be.

7. Clausal ellipsis

Another kind of regular ellipsis is the dropping of a subordinate clause that is normally required by a verb – a 'sentential complement', in the terminology of generative linguistics. We saw in example 17 that the direct object of the verb ***decline*** may be dropped in contexts where the meaning is clear. Another normal pattern of this verb is that it takes a *to*-infinitive instead of a direct object and this, too, can be dropped in appropriate circumstances, as in 25.

25. 'Take your clothes off whenever you want to,' suggests the doorman. 'You'll feel more comfortable that way.' Sarah **declines**, and we head downstairs.

The meaning, of course, is that Sarah declined (= refused) to take her clothes off. The *to*-infinitive has been elided, no doubt on grounds of obviousness.

Another example of clausal ellipsis is 26, where the self-evident clausal complement 'to steal the thing displayed' is not explicitly realized.

26. Never display anything that may **tempt** a thief.

8. Ellipsis as exploitation of a norm

In ordinary discourse, writers and speakers often omit a word when it is obvious what word or semantic type is intended (an alternation). So far in this paper, I have adduced examples illustrating the conditions under which this alternation is possible, and I have shown that it is not always possible. However, it is too early to propose a generalized account of the exploitation rule that governs the phenomenon. More corpus-driven research into the phenomenon is needed.

In other cases, 27 for example, the omission can affect the focus or the meaning of the whole sentence or, indeed, the whole discourse.

27. Stirling divided them up into eight patrols of three jeeps each, with orders to keep up the pressure. He then returned to Eighth Army Headquarters,

> accompanied by Mike Sadler. A-Squadron certainly did keep up the pressure and achieved the desired result, mining and ambushing merrily.
>
> (BNC) Anthony Kemp, 1991. *The SAS at war 1941-1945*

Normal use of the verb ***ambush*** requires a direct object. The effect of omitting it here is to suggest that it does not matter who or what was ambushed – obviously, it was the enemy. By omitting the direct object, Kemp focuses on the act of ambushing, not on the victims of the action. Who was ambushed? It does not matter. Maybe it was enemy infantry, columns of enemy tanks – or anyone or anything that happened to come along. Whoever and whatever they were, A-Squadron ambushed them. This interpretation is reinforced by the adverb ***merrily***, which would normally be regarded as inappropriate in the context of warfare. Ambushes in wartime are very far from 'merry' events: they involve fighting, destruction, and death. The effect of *merrily* is to suggest heroic nonchalance on the part of this particular group of soldiers, as they went about their business of dealing out death and destruction and risking death themselves.

The reason for classifying 27 as an exploitation rather than as an alternation is based on relative frequency and (lack of) conformity to a basic syntagmatic norm. Even in a comparatively small corpus of 100 million tokens (the BNC), several examples each of ***cultivate, frighten***, and ***punish*** dropping their direct objects can be found. On the other hand, I found no other examples of ***ambush*** dropping its direct object. Insofar as an exploitation is rare but successful, the rhetorical effect is stronger. It is not yet clear what the conditions are that permit some abnormal uses to be rhetorically effective, while others are simply mistakes or ungrammatical. It seems that the details need to be worked out word by word: an immense and daunting task.

The distinction between alternation and exploitation is mainly one of frequency. If omission of a particular argument is a regular occurrence with a given verb, it is an alternation, especially if there is little or no discernible effect on the meaning of the clause as a whole. On the other hand, if, as in 27, the omission is unusual and has a discernible effect on the interpretation, it is an exploitation of the norm.

9. Non-obvious ellipsis

Now, let us return to example sentence 3, on which I promised further comment. For convenience, it is repeated here as 28.

28. I hazarded various Stuartesque destinations like Florida, Bali, Crete and Western Turkey.

It is not immediately obvious that 28 is a case of ellipsis – but it is. Some readers – especially computational linguists and other people with a logical orientation – coming to this sentence out of context judge it to be crazy, meaningless, unidiomatic, ill-formed, or uninterpretable. But this fact merely underlines the unnatural nature of what linguists and logicians do in general and what corpus linguists do in particular.

No normal reader takes a sentence from the middle of a text and pores over it, without reference to what has gone before. Texts have a beginning, a middle, and an end. Example 28 comes from Julian Barnes' 1991 novel *Talking it over*. Barnes is a writer admired for his stylistic elegance – *The Complete Review*, for example, in a review of this novel called him "a very fine stylist" – so any problems with interpreting this sentence are unlikely to be due to infelicity or ignorance of the language on the part of the writer. In fact, when the sentence is put back into context, it makes unremarkable good sense, in a way that can only be explained in terms of exploitations of norms. The extended context is given in 28a.

> 28a. Stuart needlessly scraped a fetid plastic comb over his cranium. 'Where are you going? You know, just in case I need to get in touch.' 'State secret. Even Gillie doesn't know. Just told her to take light clothes.' He was still smirking, so I presumed that some juvenile guessing game was required of me. I **hazarded** various Stuartesque destinations like Florida, Bali, Crete and Western Turkey, each of which was greeted by a smug nod of negativity. I essayed all the Disneylands of the world and a selection of tarmacked spice islands; I patronised him with Marbella, applauded him with Zanzibar, tried aiming straight with Santorini. I got nowhere.

Various kinds of linguistic exploitation are present here. The one we are interested in is ellipsis. "I hazarded various Stuartesque destinations" is elliptical for "I hazarded a guess at various Stuartesque destinations." Having just mentioned "some juvenile guessing game", the writer does not need to repeat the word guess. A similar exploitation occurs in five subsequent clauses, in each of which a noun denoting a location or type of location (Disneylands, spice islands, Marbella, Zanzibar, Santorini) is – in its particular context – elliptical for a speech act referring to a location. Finally, there is an exploitation of considerable complexity: you aim a gun straight at something, you aim – or fire – a question at someone; you don't aim at a destination. It is noteworthy, however, that, once the scenario has been set up, these stylistic complexities do not distract from the comprehensibility of the text. No ordinary human reader puzzles over what was being hazarded, essayed, or aimed at. You either get it or you don't, but in either case an ordinary reader moves swiftly on.

10. Conclusion: norms and exploitations

The interpretation of data offered here is in line with the work of continental European dependency grammarians of the 1970s, for example Wilhelm Bondzio (1977, 1978), who proposed the notion of 'logical valency' to cover, among other things, what I am here calling elliptical arguments. It is also, I believe, compatible with frame semantics, in particular the famous paper by Fillmore and Atkins (1992), which demonstrates that, for a proper understanding of the concept **risk**, it is necessary to take account of at least five frame elements – the agent, the action, the valued object that is put at risk,

a possible bad outcome of the action, and the goal of a desired benefit. No single sentence is ever uttered containing all five elements.

This paper has attempted to show that one of the many problems confronting tasks such as phraseological analysis and mapping word meaning systematically onto word use is the implied presence of arguments that are not explicitly realized in text. If all these implied arguments of a verb were always made explicit, texts would become hopelessly overloaded and indeed unreadable. However, one contribution that a pattern dictionary can make to the interpretation of words in text is to go at least some way towards realizing explicitly the implicit arguments in patterns of word use.

Ellipsis is only one of several kinds of rule-governed linguistic exploitations of normal phraseology in natural language. The details remain to be more fully researched through detailed analysis of corpus data.

Acknowledgement

This work was supported by grant number MSM 0021620838 of the Ministry of Education of the Czech Republic.

References

BONDZIO, W. (1976/1977/1978). Abriss der semantischen Valenztheorie als Grundlage der Syntax. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 4/1976, 3/1977 und 1/1978.

CHURCH K.W. and HANKS, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1).

FILLMORE, C.J. (1975). An alternative to checklist theories of meaning. In *Papers from the First Annual Meeting of the Berkeley Linguistics Society.*

FILLMORE, C.J. and ATKINS, B.T. (1992). Towards a frame-based lexicon: the semantics of *RISK* and its neighbors. In A. Lehrer and E. Kittay (eds). *Frames, Fields and Contrasts***.** Hillsdale, New Jersey: Lawrence Erlbaum Associates.

GOLDBERG, A.E. (1995). *Constructions. A Construction Grammar Approach to Argument Structure*. Chicago: Chicago University Press.

GOLDBERG, A. E. (2006). *Constructions at Work. The Nature of Generalization in Language*. Oxford: Oxford University Press.

HANKS, P. (1994). Linguistic norms and pragmatic exploitations, or why lexicographers need prototype theory and vice versa. In F. Kiefer, G. Kiss and J. Pajzs (eds). *Papers in Computational Lexicography. Complex '94*. Budapest: Hungarian Academy of Sciences.

HANKS, P. (2008). Why the "Word Sense Disambiguation Problem" can't be solved, and what should be done instead. In B. Lewandowska-Tomaszczyk (ed.). *Corpus Linguistics, Computer Tools, and Applications. State of the Art*. Frankfurt/M.: Peter Lang.

HANKS, P. (in press). *Lexical Analysis. Norms and Exploitations*. Cambridge, MA: MIT Press.

HANKS, P., and PUSTEJOVSKY, J. (2005). A pattern dictionary for natural language processing. *Revue Française de Linguistique Appliquée*, 10(2).

HOEY, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

HUNSTON, S., and FRANCIS, G. (2000). *Pattern Grammar*. Amsterdam/Philadelphia: John Benjamins.

IDE, N. and WILKS, Y. (2006). Making sense about sense. In E. Agirre and P. Edmonds (eds). *Word Sense Disambiguation. Algorithms and Applications.* Springer.

KILGARRIFF, A., RYCHLÝ, P., SMRŽ, P. and TUGWELL, D. (2004). The Sketch Engine. In *Proceedings of Euralex 2004.* Lorient, France: Université de Bretagne Sud.

MEL'ČUK, I.A. (1988). Semantic description of lexical units in an explanatory combinatorial dictionary. *International Journal of Lexicography*, 1/3.

NIRENBURG, S. (2007). Homer, the Author of the *Iliad*, and the Computational-Linguistic Turn. In K. Ahmad, C. Brewster, and M. Stevenson (eds). *Words and Intelligence II. Essays in Honour of Yorick Wilks.* Dordrecht: Springer.

PUSTEJOVSKY, J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.

SINCLAIR, J.M. (1966). Beginning the study of lexis. In C.E. Bazell *et al.* (eds). *In Memory of J.R. Firth*. London: Longman.

SINCLAIR, J.M. (1984). Naturalness in language. In J. Aarts and W. Meijs (eds), *Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research.* Amsterdam: Rodopi.

SINCLAIR, J.M. (1987). The nature of the evidence. In J.M. Sinclair (ed.). *Looking Up: an Account of the Cobuild Project in Lexical Computing*. London and Glasgow: HarperCollins.

SINCLAIR, J.M. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Valency information online – research and pedagogic reference tools

Thomas Herbst[1], Peter Uhrig[1]
Friedrich-Alexander-Universität Erlangen-Nürnberg

Abstract

Since valency (or complementation) is an important source of errors for foreign learners of English, valency patterns of verbs, adjectives and nouns are an important aspect of EFL lexicography. English learners' dictionaries have used a number of systems for including this type of information, ranging from Hornby's verb patterns in the early OALD to transparent coding systems or pattern illustrations in the current generation of these dictionaries. The paper outlines the basic principles of the valency description provided in a special valency dictionary such as the Valency Dictionary of English (VDE) and the key features of the Erlangen Valency Patternbank, an on-line research tool based on the VDE. Since the Patternbank can be regarded as a first step towards an electronic valency dictionary for learners of English, a number of changes to make the information more immediately accessible and interpretable are discussed that may seem appropriate with respect to creating a pedagogic valency e-dictionary.

Keywords: valency, dictionary, pattern, complementation, learner lexicography, pedagogic dictionaries, learners' dictionaries.

1. Valency as an error-prone area

If one imagines – as an increasing number of linguists seem to do – that a language is (at least partly based on) a network of links between words or word forms with other words or word forms and perhaps more abstract constructions, then the analysis of collocation and valency can be regarded as two focused areas of a continuum.[2] Both are extremely important with respect to foreign language learning and teaching and thus also to foreign language lexicography since they concern item-specific knowledge about individual words. Up to a point, there may be semantic reasons to explain why certain combinations can occur or not, but there was no way of "knowing" that the

[1] {thomas.herbst,peter.uhrig}@angl.phil.uni-erlangen.de

[2] This would mean expanding Sinclair's (2004: 133) statement that "many, if not most, meanings require the presence of more than one word for their formal realization" to combinations of words and particular valency or complementation patterns. This is in line with the treatment of valency phenomena as phrasemes, *cf.* Granger and Paquot (2008: 43).

following utterances – taken from the International Corpus of Learner English (Granger *et al.* 2009) – are unacceptable for the learners who produced them.[3]

(1) ... if we *want* that also our children or our nephews will see a blue sky (and not only a grey one!) … (ICLE-IT-BER-0001.2)

(2) Just because some students *manage* getting more than three distinctions does not mean "One can, all can" (ICLE-TS-NOUN-0462.1)

(3) Man has *succeeded* to come out of the Dark Ages due to his imagination and dreams; (ICLE-BG-SUN-0076.1)

2. Coverage of valency in works of reference

That valency is a lexico-grammatical phenomenon also shows in the way it is covered in works of reference. Grammars such as the *Comprehensive Grammar of the English Language* (1985) by Quirk, Greenbaum, Leech and Svartik cover complementation phenomena at a first level by classifying verbs into classes such as intransitive, monotransitive or ditransitive, which are based on traditional categories such as direct or indirect object. Apart from the fact that such categories often involve a considerable amount of gradience because the prototypes are based on semantic and formal criteria which do not always coincide (Herbst and Schüller 2008: 167-172)[4], they are by far too general to serve learners' needs in language production. Secondly, the sections that contain the formal patterns in the kind of detail required usually contain long lists of examples of words occurring in a particular pattern, which is an indication of the fact that valency phenomena cannot be covered entirely satisfactorily by general rules but entail an element of item-specificity which also requires a lexicographical treatment.

Thus it is not surprising that the representation of complementation should have received considerable attention in lexicography: for a number of languages there are specialized dictionaries that devote themselves to the area of valency. Such valency dictionaries as, for example, the *Wörterbuch zur Distribution und Valenz deutscher Verben* (1969) by Helbig and Schenkel, the French valency dictionary by Busse and Dubost ([2]1983), *VALBU – Valenzwörterbuch deutscher Verben* (2004) or the *Valency Dictionary of English* (VDE) (2004) were explicitly created in a foreign language teaching context.

Furthermore, and probably more importantly, it is one of the outstanding characteristics of learners' dictionaries that they contain information on

[3] Such errors can also be found with nouns (*ability at flying*) or adjectives (*capable to find*), *cf.* Herbst (1985). The unpredictability of valency patterns can be illustrated by the fact that semantically similar verbs very often can be used to express the same semantic relations but do so in quite different ways. Thus a participant pattern such as AGENT – AFFECTED – PREDICATIVE can be realised by a number of formal valency patterns but not all of these valency patterns occur with all verbs showing this participant structure; *cf.* Herbst (2009a, 2009b) and Schüller (forthcoming).

[4] See also Meyer (2009).

complementation in a very systematic form: while in the first editions of the *Oxford Advanced Learner's Dictionary* a pattern system similar to that of Hornby's (1954) *Guide to Patterns and Usage in English* was used, the *Longman Dictionary of Contemporary English* ([1]1978) introduced a coding system which was designed on mnemotechnic principles. The current editions of the English learners' dictionaries either make use of transparent coding systems based on simple grammatical categories (V N to INF in OALD7) or of verbalized pattern illustrations (*want sb to do sth* in LDOCE4 or *want sb/sth to do sth* in MEDAL2). Although the latter may at first sight appear to be the more user-friendly approach, it could be argued that the use of the dynamic verb *do* actually makes it more difficult to realize that the MEDAL pattern illustration also covers sentences such as *But we want them to be accessible ...* (BNC) or *Most people do not want other people to know that they are using a dictionary* (said by Hilary Nesi in her plenary at the eLex2009 conference) than is the case with the abstract grammatical codes used in OALD7, for example – but it may well depend on the individual users and their educational background which system they prefer.[5]

3. Valency dictionaries and learners' dictionaries

It is obvious that specialized valency dictionaries differ in design and purpose from the coverage of valency patterns in general learners' dictionaries. The main aim of including valency information in a general learners' dictionary is to make it a useful tool for language production. As far as coverage is concerned, it is sufficient for the most frequent or at least established patterns for expressing a particular meaning to be listed, but it is of prime importance that it should be presented in a way that is easy for users to understand. In specialized valency dictionaries the aspect of user-friendliness is perhaps not quite so important because greater familiarity with, for example, grammatical terminology and categories of linguistic description can be expected of the users of such dictionaries. On the other hand, valency dictionaries provide much more sophisticated descriptions of valency phenomena, partly because they are based on a linguistic model that focuses on complementation and partly because they are specialized dictionaries.

In fact, most differences between the treatment of complementation in general learners' dictionaries and that in a valency dictionary such as VDE are due to the fact that a specialized dictionary is not subject to the same constraints of space as general dictionaries are. There exists one major difference in linguistic analysis, however, where the valency approach results in a different kind of classification from that provided in grammars such as CGEL: valency theory takes a more lexical perspective

[5] For a more detailed analysis of the verb pattern systems used in different editions of English learners' dictionaries see Aarts (1999), Herbst and Klotz (2003) and Herbst (2009b, 2009c). See also Klotz (2001). Note that the more recent versions of the coding systems used in learners' dictionaries do not use the functional category O (for *object*) any more. Compare, however, the survey carried out by Dziemianko (2006: 148-149). For a detailed account of the history of learners' dictionaries see Cowie (2009).

of verb complementation and regards all of the underlined elements in (4a-c) as complements of one verb *decide*:

(4a)$_{VDE}$ He *decided* on roast chicken and vegetables

(4b)$_{BNC}$ She *decided* to take a look at the harbour.

(4c)$_{BNC}$ She *had decided* that she wanted to go to London.

Grammars such as CGEL tend to analyse cases such as (4a) as an instance of a category prepositional verb, which can then be seen as fitting into a monotransitive pattern of some kind.[6] The concept of prepositional verbs has a detrimental effect when applied to lexicography since classifying *decide on* as a prepositional verb often means that complementation patterns of *decide* with *on*-phrases are given in a different place from those with a *to*-infinitive or a *that*-clause.[7] Furthermore, dictionaries such as MEDAL2 treat cases such as *decide on* or *refer to* in the same way as combinations with shiftable particles such as *look up*, although only the latter are referred to as phrasal verbs in CGEL. Interestingly, *succeed in* is not categorised as a phrasal verb in LDOCE5, OALD7 or MEDAL2. Treating elements such as *on roast chicken and vegetables* in (4a) as a particular type of complement, which can occur with verbs, nouns or adjectives, not only results in a theoretically more adequate account of valency phenomena but also in greater user-friendliness.

A detailed valency description also entails a specification of complements as obligatory, contextually-optional or optional. Thus the fact that a sentence such as

(5)$_{VDE}$ How could I possibly forget?

can only be used in a situation in which the referent of the second participant of *forget* can be identified is made clear in the VDE by marking the corresponding valency slot as "cont(extually optional)" and adding a label "only if clear from context" to the examples – which in this explicitness could not be done systematically in general learners' dictionaries.

The same holds for an indication of the meanings of different valency patterns and the semantic and/or lexical restrictions for particular complements. Although it could be argued that "ideally" a valency description should specify each valency slot semantically in terms of general semantic roles or more specific participant roles

[6] Interestingly, Quirk, Greenbaum, Leech and Svartvik (CGEL 16.5) classify such cases as prepositional verbs type I: according to CGEL (16.5), noun phrases such as *roast chicken* in (4a) would be described as the complement of the preposition and not as the object of a prepositional verb *decide on*; in CGEL (16.28) *decide on* is included under the heading of "variants of monotransitive complementation".

[7] There is a considerable amount of inconsistency to be observed: for example, in LDOCE5 *decide on/upon* is given as a subentry of *decide*, but *decide against* and *decide in favour of* are included under *decide*. Note that the rejection of the notion of prepositional verbs within a valency framework only concerns prepositional verbs in the sense as defined in CGEL and not phrasal verbs. Thus cases such as *decide + on* NP are distinguished in VDE from idiomatic phrasal verbs such as *look up*, which are treated as multi-word lexical units.

(Herbst and Schüller 2008: 126-134) or in terms of frame elements (Fillmore 2007), such roles are only of limited value for lexicographic purposes. VDE thus makes use of a more flexible way of describing such semantic – and also collocational – properties of patterns in the form of notes like the following (2004: 585):

> A person or something such as an argument or a fact can persuade a person
>
> (i) **to do something**, *i.e.* make them do it.
>
> (ii) **that something is the case**, *i.e.* make them believe that it is true.
>
> (iii) **into or out of something or doing something**, *i.e.* make them do or stop doing it.
>
> (iv) **of the need, advantage, benefit, etc. of sth.**, *i.e.* make them believe in it.

A further fundamental difference between general learners' dictionaries and valency dictionaries is that the latter should aim at comprehensive coverage of all the valency patterns that can be found for a particular lemma – if possible, with an indication of their frequency. This is particularly important with respect to the function of valency dictionaries as research dictionaries that can be used for developing pedagogic material and as marking dictionaries for non-native teachers of the language, which have to contain rare patterns so that they do not get marked wrong and to indicate highly frequent patterns whose use ought to be encouraged.[8]

The character of specialized dictionaries has important repercussions as to their overall purpose: while a dictionary such as VDE can certainly be used as a production dictionary by learners, it would be unrealistic to see this as its prime purpose – given its price, size and the relatively small number of lexical items covered. However, the clash arising from this conflict between the lexicoparameters of user-friendliness and depth of description is particularly apparent in the case of traditional print dictionaries – simply because users cannot easily be made to ignore information that is irrelevant to their current needs.[9]

It is obvious that the electronic media open up new possibilities in this respect in that the same database can be used to serve different kinds of "concrete"[10] user needs. The consequence must be to develop flexible access structures along the principles realized for example by Bergenholtz and his colleagues in *Ordbogen over Faste Vendinger*, where users are asked to specify their needs with respect to such parameters as decoding or encoding and are accordingly provided with different types of information – in other words, what is needed is customized access to the information provided in a valency dictionary.

[8] Achieving this aim is subject to obvious restrictions caused by factors such as limitations of the corpora used, oversight and the fact that the complement/adjunct distinction can be subject to gradience.

[9] Compare Engelberg and Lemnitzer (2008: 224): "Einzelne Segmente der Mikrostruktur von Wörterbuchartikeln sollten ausgeblendet werden können". ["It should be possible to hide specific elements of dictionary entries."]

[10] For the concreteness of user needs and different types of user needs see Tarp (2009: 46-48).

4. The Erlangen Valency Patternbank

The options opened up by the electronic medium go far beyond those related to offering customized types of information since they also allow different access routes to the data presented. A first attempt at providing online access to the VDE data is presented by the Erlangen Valency Patternbank, which contains the patterns of the 511 verbs, 544 adjectives and 274 nouns covered in the VDE.[11] Although the number of lemmata may seem rather small, we estimate that the verbs contained in the VDE can account for more than two thirds of all verb uses in the BNC, which means that token coverage will not increase dramatically by including new verbs, which we are nevertheless planning to do.

The Patternbank has been designed entirely for research purposes. Its main advantage is that it provides a completely new access structure, some features of which will be outlined below. Users can choose between searches providing

- lists of all active and passive verb patterns, adjective patterns and noun patterns occurring with the lemmata of the VDE (search for pattern – see Figure 1),
- lists of all patterns containing a particular complement (search for pattern element) and
- lists of all patterns occurring with a particular word (lexical search).

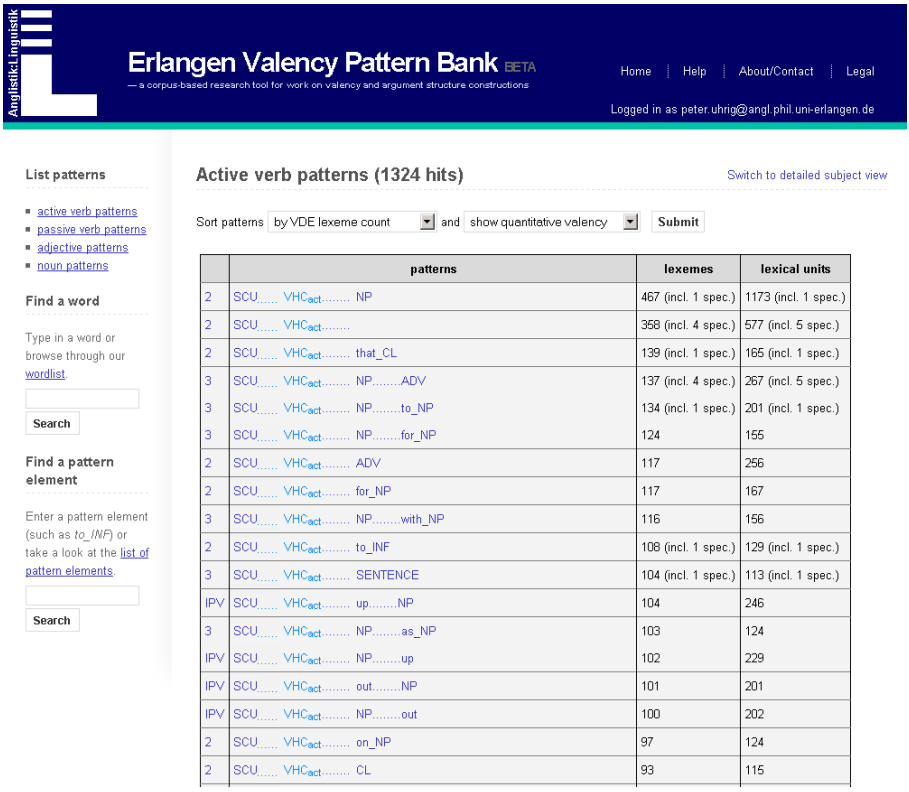
The patterns are presented in a very similar form as in VDE, in terms of a formal specification of the complements in terms of the type of phrase or clause realizing a particular valency slot, *i.e.* by symbols such as [NP], [that_CL] or [to_INF]. There are two exceptions to this: in verb patterns, a distinction is made between patterns taking impersonal [it] or impersonal [there] as subjects and the code SCU, which stands for subject complement units that can be realized by a noun phrase or possibly further elements.[12] The verbal head complex – VHC –, *i.e.* that part of the predicate that contains the governing verb and auxiliaries preceding it, is specified with respect to active and passive voice.[13] The patterns can be accessed according to different sorting principles:

- arranged in terms of the number of lexemes or lexical items in which the patterns occur in the Patternbank,
- strictly alphabetically or
- alphabetically according to the first complement after the verb.

[11] For copyright reasons, the present version of the Patternbank does not include the examples and the notes on meaning given in the VDE. In the case of polysemous words, however, the Patternbank includes an indication of the corresponding sense in VDE. Furthermore, a mouseover system shows the corresponding pattern numbers in VDE to facilitate the search for examples in the dictionary.

[12] More detailed information of the possibility of non-NP subjects can be retrieved in the "detailed subject view", if required. It has to be pointed out, however, that this type of information relies largely on native speaker intuition in VDE.

[13] See Herbst and Schüller (2008) for a more detailed account of this.



	patterns	lexemes	lexical units
2	SCU..... VHCact..... NP	467 (incl. 1 spec.)	1173 (incl. 1 spec.)
2	SCU..... VHCact.....	358 (incl. 4 spec.)	577 (incl. 5 spec.)
2	SCU..... VHCact..... that_CL	139 (incl. 1 spec.)	165 (incl. 1 spec.)
3	SCU..... VHCact..... NP.....ADV	137 (incl. 4 spec.)	267 (incl. 5 spec.)
3	SCU..... VHCact..... NP.....to_NP	134 (incl. 1 spec.)	201 (incl. 1 spec.)
3	SCU..... VHCact..... NP.....for_NP	124	155
2	SCU..... VHCact..... ADV	117	256
2	SCU..... VHCact..... for_NP	117	167
3	SCU..... VHCact..... NP.....with_NP	116	156
2	SCU..... VHCact..... to_INF	108 (incl. 1 spec.)	129 (incl. 1 spec.)
3	SCU..... VHCact..... SENTENCE	104 (incl. 1 spec.)	113 (incl. 1 spec.)
IPV	SCU..... VHCact..... up.....NP	104	246
3	SCU..... VHCact..... NP.....as_NP	103	124
IPV	SCU..... VHCact..... NP.....up	102	229
IPV	SCU..... VHCact..... out.....NP	101	201
IPV	SCU..... VHCact..... NP.....out	100	202
2	SCU..... VHCact..... on_NP	97	124
2	SCU..... VHCact..... CL	93	115

Figure 1. Top of the list of active verb patterns in the Erlangen Valency Patternbank.

The Patternbank then provides a variety of different viewing options:

- Clicking on a particular pattern yields a list of all lexical units occurring in this pattern – verbs that take a corresponding passive pattern are marked in colour.
- Clicking on a word in such a list leads to an inventory of all patterns of this word, thus allowing a lexical perspective (which can be accessed directly in a lexical search).

The Patternbank tries to provide a description of valency patterns that is not restricted to one particular theory of language and thus aims to be as neutral in its descriptive categories as possible. This is also reflected in the presentation of cases where there might be differences of opinion as to the exact pattern status. For this reason, the notions of lexically specified patterns was introduced to cover cases such as *drop dead*, which can either be seen as instances of verb + adjective phrase or as idiomatic combinations. In a similar way, the category of contextually specified patterns contains elements such as *at a loss what to do*, where the wh_to_INF can be seen as a pattern of *at a loss* rather than the noun *loss* as such. In the long run, it might be more appropriate to treat such cases in terms of multi-word valency carriers.

Through the various access modes the Patternbank opens up new perspectives for research: thus it becomes relatively easy to compare which patterns occur with the same or a similar range of lexical items, which is an important instrument for linguistic research and in fact also for discovering inconsistencies in pattern coverage in our own data. The lists provided by the Patternbank can also serve to carry out research of a more theoretical interest such as whether all instances of a valency pattern can also be seen as expressing the same semantic roles, *i.e.* as representing the same valency construction (Herbst and Schüller 2008), which could be important with respect to the status of argument structure constructions in certain frameworks of construction grammar (Goldberg 2006).[14] By making the Patternbank freely available – being fully aware of shortcomings concerning completeness of the data and a number of classificatory problems – we hope to provide a tool for a range of purposes in theoretical and applied linguistics.[15]

5. Towards a pedagogic valency dictionary

Apart from being a research tool for linguists, the Erlangen Valency Patternbank can be seen as a first step towards an electronic valency dictionary. As pointed out above, such a dictionary should be designed in such a way that the same database can be used to serve three main functions:

- language production (quite obviously for learners at different levels)
- marking (for non-native teachers of the language)
- research (for empirical research in theoretical linguistics or applied linguistic purposes)

With the exception of the number of patterns shown – where there is an important difference between what is required for language production on the one hand and the other two purposes on the other – the type of information required seems to be very much the same for all three of these purposes: targeting different user groups and different user needs is thus not so much a question of which information to provide but in which form and in which order this information should be provided.[16]

Customization of information could concern a number of aspects of VDE-online. For example, customized electronic access may contribute to overcoming the clumsiness of a printed VDE entry, where some users may find the cross-reference system between patterns, examples and meaning notes tiresome.

[14] For such uses of the Patternbank see Herbst (2009d).

[15] The Patternbank can be accessed at www.patternbank.uni-erlangen.de, where a feedback function can be used to send in suggestions for improvement.

[16] For certain user groups, one might consider including an option containing simplified corpus-based examples that can easily be related to the corresponding patterns.

Language learners using the database for production purposes are likely to look for uses of a particular word. They should then be able to choose the appropriate meaning of that word and be given a list of established ways of expressing a particular meaning. Particularly frequent patterns could be highlighted to encourage their use; rare patterns will not be indicated at all – provided there are more established patterns expressing that particular meaning; otherwise they will be shown but marked as "rare".

It would be highly desirable for production purposes to have a search option that not only provides the established valency patterns of the lexical unit the user has asked for but also shows further ways of expressing similar meanings involving other headwords – *i.e.* to have a thesaurus-like search tool providing for example the patterns of *resemble, similar* as well as *resemblance*. One could also imagine searches based on collocates or meaning elements.

In any case, in the production mode it is also necessary that examples be shown immediately after the pattern and that the meaning notes of VDE, which provide information on semantic roles and collocational restrictions, could be made to appear at the same time.

For the language teacher wanting to check whether a particular valency pattern exists at all, the starting point of the search is the pattern list. If one wants to find out whether a particular verb occurs in a particular pattern, either the verb list given for the pattern or the pattern list given for the verb can be used. If [to_INF] does not occur in the list of active patterns – as it would not for *succeed*, for example – then ideally this can be taken as an indication of the fact that this pattern does not exist; only if it does occur are examples and meaning notes needed at all.[17]

Customization could also affect the presentation of the patterns as such. While linguists may see the value of theory-related symbols such SCU or VHC and be happy to familiarize themselves with terms such as "contextually-optional complement", many users are bound to prefer more established and simpler terms. Again, one could imagine offering different ways of pattern representation – but it is important to remember that there are no easy solutions. For example, experience has shown that a symbol such as N_P (marking noun phrases that can function as the subject of passives in VDE) is easily misinterpreted to stand for noun phrase. One solution is to use more explicit subscripts such as $N_{pass\text{-}subj}$, another to list active and passive verb patterns separately. This, however, results in a greater number of patterns, which may be equally confusing for users. Furthermore, listing active and passive patterns separately causes the lexicographical problem that it may not always be possible to find suitable examples of every passive use in the corpus.

[17] Strictly speaking, the fact that an active pattern is not listed in the Patternbank for a particular verb only means that it was not identified as such in the process of the compilation of the VDE. Information on passive patterns has to be treated with greater caution since the VDE descriptions on which the Patternbank relies at the moment are mostly based on active patterns.

What may be even more important is the question of whether N or NP should be used to represent the category of noun phrase. From a linguistic point of view, NP may seem more adequate – in fact, it could be argued that the importance of the level of the phrase is regrettably underrated in many textbooks, but that is also why many users may find NP disturbing rather than helpful. The alternative chosen in learners' dictionaries such as LDOCE and MEDAL, to use pattern illustrations of the *want to do sth* type, has the obvious disadvantages of being potentially misleading semantically and of not being easily applicable to the more complex patterns to be covered in a valency dictionary. Nevertheless, this could be an option for the production mode for not very advanced learners. In any case, the question of symbols becomes less problematic in the electronic medium if a mouse-over function is installed that contains an explicit explanation of the symbols used to describe a pattern.

6. Conclusion

It is obvious that the possibilities created by e-lexicography are enormous, also as far as the representation of valency information is concerned. An online valency dictionary could overcome many of the weaknesses of printed valency dictionaries and bridge the gap to learners' dictionaries – either as a separate electronic tool or as one component of an electronic lexicographical and grammatical information device.[18] The main advantage of the medium is that it makes it possible to provide information that serves users' needs more directly and more immediately than is possible in printed dictionaries by interactive access structures allowing a high degree of customization. In order not to throw out the baby with the electronic bathwater it is important to remember that customization only makes sense if there are customers – customers in the sense of language learners and language teachers who are aware that the product exists, how to use it and that it serves their needs, *i.e.* users who know what their needs are. Fascinating electronic research tools that provide users easy access to valency patterns, collocations and other features of language will only be a success if there is a sufficiently large number of people around who understand the idiosyncratic and collocational aspects of language for which Sinclair (1991, 2004) introduced the term *idiom principle*. The development of such tools must thus go hand in hand with preparing the ground for their use in foreign language teaching, which means that we have got to make sure that "[c]oncepts such as word grammar, colligations, collostructions and patterns" in future do "rank very highly on teachers' priority lists" (Granger and Meunier 2008: 248).

[18] This could take the form of a plug-in feature for electronic learners' dictionaries (similar to the native-language add-ons suggested by Leech and Nesi (1999: 300)), which could even be designed to allow the authors of the plug-in to override the valency information in the existing dictionary.

References

AARTS F. (1999). Syntactic information in OALD5, LDOCE3, COBUILD2 and CIDE. In T. Herbst and K. Popp (eds). *The Perfect Learners' Dictionary(?)* Tübingen: Niemeyer: 15-32.

COWIE, A. (2009). The earliest foreign learners' dictionaries. In A. Cowie (ed.). *The Oxford History of English Lexicography. Volume II Specialized Dictionaries*. Oxford: Clarendon Press: 385-411.

DZIEMIANKO, A. (2006). *User-Friendliness of Verb Syntax in Pedagogical Dictionaries of English*. Tübingen: Niemeyer.

ENGELBERG, S. and LEMNITZER, L. ([3]2008). *Lexikographie und Wörterbuchbenutzung*. Tübingen: Stauffenburg.

FILLMORE, C. (2007). Valency issues in FrameNet. In T. Herbst and K. Götz-Votteler (eds). *Valency. Theoretical, Descriptive and Cognitive Issues*. Berlin/New York: Mouton de Gruyter: 129-160.

GOLDBERG A. (2006). *Constructions at Work. The Nature of Generalizations in Language*. Oxford/New York: Oxford University Press.

GRANGER S., DAGNEAUX, E., MEUNIER, F. and PAQUOT, M. (2009). *International Corpus of Learner English v2*. Louvain-la-Neuve: Presses universitaires de Louvain.

GRANGER S. and PAQUOT, M. (2008). Disentangling the phraseological web. In S. Granger and Fanny Meunier (eds). *Phraseology. An Interdisciplinary Perspective*. Amsterdam/ Philadelphia: Benjamins: 27-50.

GRANGER, S. and MEUNIER, F. (2008). Phraseology in language learning and teaching. Where to from here? In F. Meunier and S. Granger (eds). *Phraseology in Foreign Language Learning and Teaching*. Amsterdam/Philadelphia: Benjamins: 247-252.

HERBST, T. (1985). Von Fehlern, die vermeidbar wären. Ein weiteres Argument für mehr Wörterbucharbeit im Englischunterricht. In E. Zöfgen (ed.). *Wörterbücher und ihre Didaktik. Bielefelder Beiträge zur Sprachlehrforschung*, 14.1/2: 236-248.

HERBST, T. (2009a). Valency. Item-specificity and idiom principle. In U. Römer and R. Schulze (eds). *Exploring the Lexis-Grammar Interface*. Amsterdam/Philadelphia: Benjamins: 49-68.

HERBST, T. (2009b). Patterns in syntax, lexicography and corpus linguistics. In L. Eckstein and C. Reinfandt (eds). *Anglistentag 2008 Proceedings*. Trier: WVT: 379-289.

HERBST, T. (2009c). Item-specific syntagmatic relations in dictionaries. In S. Nielsen and S. Tarp (eds). *Lexicography in the 21st Century*. Amsterdam/Philadelphia: Benjamins: 281-308.

HERBST, T. (2009d). Introduction. http://www.patternbank.uni-erlangen.de/cgi-bin/ patternbank. cgi?do=introtxt.

HERBST, T. and KLOTZ, M. (2003). *Lexikografie*. Paderborn: Schöningh (UTB).

HERBST, T. and SCHÜLLER, S. (2008). *Introduction to Syntactic Analysis. A Valency Approach*. Tübingen: Narr.

HORNBY, A.S. (1954). *A guide to patterns and usage in English*. London: Oxford UP.

KLOTZ, M. (2001). Valenzinformation im monolingualen englischen Lernerwörterbuch und im bilingualen Wörterbuch englisch-deutsch. *Zeitschrift für Angewandte Linguistik*: 61-79.

LEECH, G. and NESI, H. (1999). Moving towards perfection. The learners' (electronic) dictionary of the future. In T. Herbst and K. Popp (eds). *The Perfect Learners' Dictionary(?)* Tübingen: Niemeyer: 295-308.

MEYER, M. (2009). Revisiting the evidence for objects in English. In R. Schulze and U. Römer. *Exploring the Lexis-Grammar Interface*. Amsterdam/Philadelphia: Benjamins: 211-227.

QUIRK, R., GREENBAUM, S., LEECH, G. and SVARTVIK, J. (1985). *A Comprehensive Grammar of the English Language*. London/New York: Longman. [CGEL]

SINCLAIR, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

SINCLAIR, J. (2004). *Trust the Text. Language, Corpus and Discourse*. London/New York: Routledge.

SCHÜLLER, S. (forthcoming): *Semantic Aspects of Verb Valency. The Relationship between Meaning and Form* [Dissertation Erlangen].

TARP, S. (2009). Reflections on data access in lexicographic works. In S. Nielsen and S. Tarp (eds). *Lexicography in the 21st Century*. Amsterdam/Philadelphia: Benjamins: 43-62.

Dictionaries and electronic databases:

A Valency Dictionary of English. A Corpus-Based Analysis of the Complementation Patterns of English Verbs, Nouns and Adjectives (2004). T. HERBST, D. HEATH, I. ROE and D. GÖTZ (eds). Berlin/New York: Mouton de Gruyter. [VDE]

Erlangen Valency Patternbank (2009-). T. HERBST and P. UHRIG. Available online at http://www.patternbank.uni-erlangen.de

Französisches Verblexikon. Die Konstruktion der Verben im Französischen. (^{2}1983) W. BUSSE and J.-P. DUBOST. Stuttgart: Klett.

Longman Dictionary of Contemporary English (1978). P. PROCTER (ed.). Harlow: Longman. [LDOCE1]

Longman Dictionary of Contemporary English (^{4}2003). D. SUMMERS (ed.). Harlow: Longman. [LDOCE4]

Longman Dictionary of Contemporary English (^{5}2009). M. MAYOR (ed.). Harlow: Longman. [LDOCE5]

Macmillan English Dictionary for Advanced Learners (^{2}2007). M. RUNDELL (ed.). Oxford: Macmillan. [MEDAL2]

Ordbogen over Faste Vendinger. (2007). H. BERGENHOLTZ, V. VRANG and R. ALMIND. Aarhus: Centre for Lexicography, Aarhus School of Business. [www.idiomordbogen.dk]

Oxford Advanced Learner's Dictionary (^{7}2005). S. WEHMEIER (ed.). Oxford: Oxford University Press. [OALD7]

VALBU – Valenzwörterbuch deutscher Verben. (2004). H. SCHUMACHER, J. KUBCZAK, R. SCHMIDT and V. DE RUITER. Tübingen: Narr.

Wörterbuch zur Valenz und Distribution deutscher Verben (1969/^{2}1973) G. HELBIG and W. SCHENKEL (eds). Leipzig: Enzyklopädie.

Automated collection of Japanese word usage examples from a parallel and a monolingual corpus

Kristina Hmeljak Sangawa[1], Tomaž Erjavec[2], Yoshiko Kawamura[3]
University of Ljubljana, Jožef Stefan Institute, Tokyo International University

Abstract

Examples are an important source of information on word usage for language learners, but existing reference sources for Japanese as a second language are limited. This paper describes two projects for the automated collection of word usage examples. Examples extracted from an ad-hoc Japanese-Slovene parallel corpus were included into jaSlo, a Japanese-Slovene learners' dictionary, and examples extracted from a monolingual web-harvested 400 million word corpus of Japanese were selected to be used as supplementary examples for Chuta, a multilingualized dictionary for learners of Japanese as a second language.

Keywords: example, word usage, corpus example, Japanese web corpus, Japanese-Slovene parallel corpus, readability.

1. Introduction

Examples are an excellent source of semantic, syntactic, morphological, collocational and pragmatic information for dictionary users, especially for those who are not familiar with lexicographic metalanguage and prefer inferring (or guessing) word usage from examples rather than from definitions or symbols. However, although many examples can be included into electronic dictionaries where space is not as limited as in paper dictionaries, good examples, which should be typical, natural and surrounded by typical context (Fox 1987: 37, Atkins and Rundell 2008: 330), are costly to produce (Rychlý *et al*. 2008: 425). This is especially crucial in the case of voluntary-based or low-budget academic lexicographic projects with limited human and financial resources.

For learners and teachers of Japanese as a second language, examples of word usage can be found in existing dictionaries, textbooks and corpora, but each of these sources has some limitations. Starting with the *Dictionary of basic Japanese usage for foreigners* (Bunkachô 1971), a number of monolingual, bilingual and bilingualized dictionaries for learners of Japanese as a second/foreign language have been produced

[1] University of Ljubljana, kristina.hmeljak@ff.uni-lj.si
[2] Jožef Stefan Institute, tomaz.erjavec@ijs.si
[3] Tokyo International University, kawamura@tiu.ac.jp

in the last three decades, and all of them contain usage examples. However, dictionaries covering at least 10,000 headwords (*i.e*. the vocabulary considered to be needed by intermediate to advanced learners of Japanese, *cf.* Tamamura (1984), Japan Foundation (2004)) such as the *Informative Japanese dictionary* (Nihongo no kai 1995) with 11,000 headwords, or *Kodansha's Communicative English-Japanese Dictionary* (Sharpe 2006) with 22,000 entries usually do not offer more than 2-3 examples per headword. On the other hand, those containing more examples per headword such as the monolingual *Japanese dictionary: learning language the fun way* (Takano 2004) with 750 headwords, the dictionary of functional words *Nihongo bunkei jiten* (Group Jamashii 1998) with 3000 entries, the bilingual *Kodansha's Basic English-Japanese Dictionary* (Makino *et al.* 1999) with 4500 headwords, or the *Effective Japanese Usage Guide* (Hirose & Shoji 1994) with 708 headwords, do not cover all the vocabulary needed by intermediate to advanced learners of Japanese.

There is a very large number of monolingual and bilingual dictionaries for native speakers of Japanese, which also contain examples. However, Japanese large monolingual dictionaries for native speakers, such as *Kôjien (*Shinmura 2008) or *Daijirin* (Matsumura 2006), are generally too difficult for foreign learners, especially those based on historical principles, which contain examples of archaic language, such as *Kôjien.* The very numerous monolingual dictionaries for elementary-school children, such as *Reikai shôgaku kokugo jiten* (Tajika 2009), *Shôgaku shin kokugo jiten* (Kai 2002) or *Challenge shôgaku kokugo jiten* (Minato 2008), or for high-school native speakers, such as *Meikyô kokugo jiten* (Kitahara 2002) or *Shin meikai kokugo jiten* (Yamada 2005), which do contain easier examples with phonetic script, do not usually contain more than 2-3 examples per word. On the other hand, examples in bilingual dictionaries for Japanese native speakers are usually targeted at explaining and translating idiomatic examples in the foreign language. The Japanese translations in these dictionaries do not always exemplify the most typical usages of Japanese words, but rather collocations and phrases which are difficult to translate.

Another obvious source of examples is corpora. Some Japanese corpora have been made available in the last few years. A 39 million word demo version of the Balanced Corpus of Contemporary Written Japanese (BCCWJ) being compiled at the National Institute for Japanese Language (Maekawa 2008), was made available on the institute's portal in March 2009. Examples from the Japanese Web as Corpus (JpWaC), a 400 million word corpus of web text (Srdanović Erjavec *et al*. 2008), can be looked up via the Sketch Engine (Kilgarriff *et al*. 2004). Search engines such as Google or Yahoo can also be used as a source of usage examples, although the lists of search results given in such engines are neither linguistically representative nor sortable by linguistic criteria. However, results from such corpora searches can be overwhelming for language learners with limited linguistic ability.

We therefore decided to create an intermediate tool which would give the users (especially language learners) more examples than a conventional dictionary or textbook, but which would be less overwhelming than corpora or search engine results.

Having limited human and financial resources, we tried to make the best possible use of available resources.

In the following sections we present two projects where usage examples were automatically collected to be included in two electronic dictionaries for learners of Japanese. Both dictionaries are being compiled as academic projects with the help of volunteer editors and progressively published on the web.

2. Examples from a parallel corpus

Bilingual usage examples were collected for jaSlo (Erjavec *et al.* 2006), a Japanese-Slovene dictionary for Slovene learners of Japanese, which is being compiled at the University of Ljubljana and has been gradually published at http://nl.ijs.si/jaslo/ since 2001. The dictionary's latest edition (2006) had c. 10,000 Japanese headwords and c. 25,000 Slovene translational equivalents, but only 2,370 usage examples. We therefore decided to use an existing collection of parallel texts, augment it and structure it into a parallel corpus to use it as a source of examples.

2.1. Parallel corpus compilation

Due to a lack of competent translators for the language pair Japanese-Slovene, hardly any translation had been produced between these two languages before the establishment of the Japanese studies program at the University of Ljubljana in 1995. However, since its establishment, a small collection of parallel texts in electronic form accumulated at the department, consisting of lecture handouts (academic texts on the history, literature, geography and society of Japan, prepared in Japanese by visiting lecturers and translated into Slovene by university staff), and student coursework (texts on the Japanese society translated from Japanese into Slovene and texts on tourism translated from Slovene to Japanese, in both cases translated by students and thoroughly revised by the teacher in charge of the translation course). Given their availability in electronic form, we decided to use them as sources of examples for our Japanese-Slovene dictionary. However, since most texts were quite challenging for language learners, we decided to add some more readable texts, from which examples for intermediate students could be obtained. We therefore digitized parts of 6 contemporary Japanese novels which were translated into Slovene in the last decade and the only novel that has been translated from Slovene into Japanese up to now, by scanning them and manually correcting OCR output.

Lastly, in order to obtain a larger corpus, we searched the web for translated pages in Japanese and Slovene. We searched for pages in Japanese script within the domain .si (Slovenia), and found 150 pages, of which only 4 were relevant translations, while the others were either brief Japanese summaries of longer Slovene texts or poor quality machine translation products. We also searched for pages written in Slovene in the .jp domain, using Google's advanced search function to limit the target language, which yielded a few hundred pages, but only two of them were found to be relevant, while all

the others were wrongly identified as Slovene and actually written in some other Slavic language. We carried out searches for the words “Japanese” and “Slovene” without specifying their internet domain and found pages, mostly in English, localized into many languages including Japanese and Slovene, where the names of the two languages appeared in a menu for language selection. Such indirect translations are certainly not ideal sources of dictionary examples, but given the lack of direct translations, we decided to include them, after manually checking them and removing all segments with unreliable or missing translations.

All texts were normalized into plain text files and sentence-aligned using Wordfast PlusTools (www.wordfast.net). Alignment was manually validated, and paragraphs with missing or mistaken translations were removed. The complete corpus was lemmatized using Chasen (Matsumoto *et al*. 2007) for the Japanese part and “totale” (Erjavec *et al*. 2005) for the Slovene part. Combining all available parallel texts, we obtained a parallel corpus of 7,914 translation units (sentence pairs), corresponding to 226,220 Japanese morphemes and 171,261 Slovene words, and composed of the following subcorpora: translated lecture handouts (13.5%), revised student translations (24.5%), literary fiction (15.7%), and multilingual web pages (46.3%).

2.2. Extraction of examples from the parallel corpus

All Japanese headwords in the dictionary were searched for in the corpus, yielding examples for 4,648 lemmas, *i.e*. approximately half the dictionary entries. When more than 6 examples were found, only the shortest 6 were retained, since short sentences tend to be syntactically simpler. Lexical complexity was not taken into account at this stage, but all sentences are accompanied by a translation into the user’s native language, and therefore presumably understandable.

Examples were appended to the dictionary entries, and graphically separated from existing examples, as can be seen in Figure 1. This version of the dictionary was published at http://nl.ijs.si/jaslo/cgi/jaslo-eg.pl.

Each corpus example is followed by a link (in the form of an arrow), which leads to a page with information on the title, place and date of publication or URL of the source text and translated text, source language and target language, author’s and translator’s name when available, thus indicating in what sort of genre the word can be found.

2.3. Evaluation of extracted examples

Automatically extracted corpus examples did not go through the usual editorial process of dictionary entries, *i.e.* analysis of a corpus of examples, synthesis of the dictionary entry and editing of appropriate examples. It cannot therefore be expected, especially given the very small size of our corpus, that automatically extracted examples should cover all senses of a word or give all its most typical syntactic and collocational patterns. We evaluated a sample of 80 lemmas of intermediate difficulty,

randomly chosen from the Japanese Language Proficiency Test specifications (Japan Foundation 2004) to test coverage of word senses and usefulness.

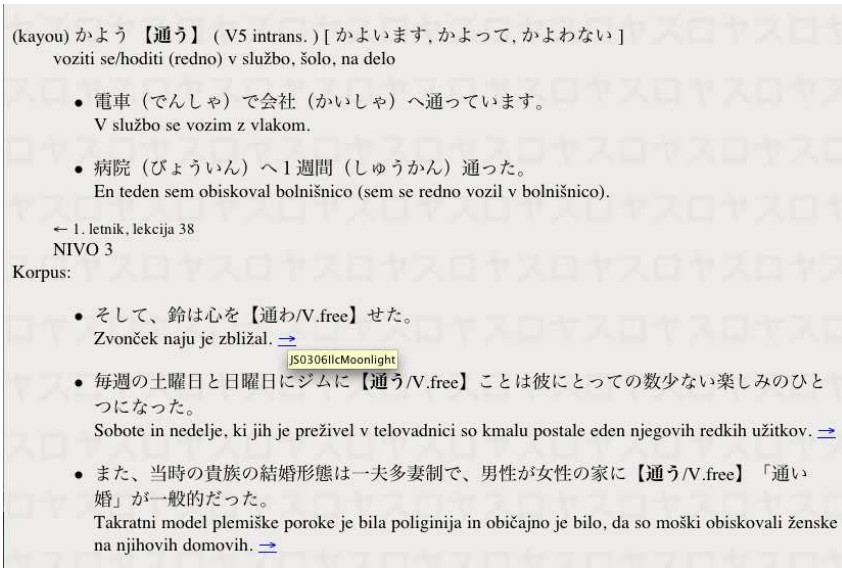


Figure 1. Example of a jaSlo dictionary entry with corpus examples

We found that for 51% of these lemmas, all senses were covered, while for the remaining half of the lemmas some senses did not appear in any of the examples. This indicates the need for a larger corpus to achieve better coverage. For 10% of the lemmas, new translational equivalents were found which had not yet been included in the latest version of the dictionary: 2% were context dependent or unnecessarily liberal translations which were not deemed useful, but as much as 8% were useful additions to the dictionary. Moreover, corpus examples for 4% of the sampled headwords contained idiomatic expressions, collocations or multi-word units which were not present in the original dictionary, and therefore useful additions to it.

Given the fact that the dictionary was compiled by a small team of contributors with little lexicographic experience and that there are few Japanese-Slovene contrastive studies or other reference materials, it is not surprising that the dictionary still needs improvement. These corpus examples are therefore going to be useful not only for the general users, but also for the editors of the dictionary during future revisions.

Regrettably, 8% of the examples were assigned to the wrong dictionary entry because of lemmatization errors in Chasen's morphological analysis. This indicates the need for a future manual validation of example selection, while for the time being users are warned that corpus examples were extracted automatically and may contain errors.

In the future, we plan to augment the parallel corpus to achieve better coverage, and to enhance the example selection procedure to include readability criteria (including

vocabulary coverage, syntactic patterns and context independence) and typicality criteria (including collocational, morphosyntactic and stylistic patterns).

3. Examples from a monolingual corpus

In a similar pilot study, a corpus of usage examples was collected to be combined with Chuta (Kawamura and Kaneniwa 2006, published at http://chuta.jp/), a multilingualized Japanese learners' dictionary in which sense divisions, definitions and usage examples are first prepared by a team of Japanese native speakers, teachers of Japanese as a second language, and subsequently translated into different languages by an international team of editors (Vietnamese, Russian, English, Turkish, Bulgarian, Korean, Chinese, Portuguese, Spanish, German, Czech, Malay, Kirghiz, Marathi, Slovak, Thai, French, Italian, Finnish, Nahuatl, Slovenian, Indonesian, Hungarian, Tagalog, Arabic and Romanian, in decreasing order of number of edited lemmas). 8,721 Japanese entries have been published at present, while bilingual entries are still being edited. To increase the number of examples for general users and also help editors of bilingual entries, examples were extracted from a web-harvested, lemmatized and PoS tagged 400 million word corpus of Japanese, JpWaC (Srdanović *et al*. 2008).

3.1. Compilation of JpWaC-L2, a monolingual corpus of example sentences

A 100 million word sample of the JpWaC corpus was extracted, starting from the beginning of the corpus until the required size was obtained. As the corpus texts are sorted according to the URL, and the start of the URL is essentially random, this does not unduly bias the corpus composition. The corpus is composed of texts, each marked by its source URL, and these, in turn, composed of sentences, each annotated by the sequential number of the sentence in the text. All words in the corpus were annotated with their difficulty level according to the Japanese Language Proficiency Test specifications, ranging from 4 (easiest words) to 1 (hardest words). Words not appearing in the JLPT list were assigned level 0. Each sentence was furthermore annotated with quantitative information for the number of tokens in the sentence, words by levels, punctuation symbols and numerals.

Single sentences were extracted from this sample corpus to create a corpus of example sentences, JpWaC-L2, according to the following criteria. We retained sentences which:

a. are not duplicate (only the first occurrence of duplicate sentences is retained);

b. are between 5 and 25 tokens in length (to exclude very short sentences, which are usually only sentence fragments, and very long sentences, which are difficult to understand);

c. contain less than 20% of punctuation marks or numerals (to retain only text rich sentences);

d. contain at most 20% of level 0 words (to exclude sentences with a high proportion of difficult or foreign words);

e. do not contain words written with non-Japanese characters (to exclude strings such as URLs, e-mail addresses or text in other languages);

f. do not contain any opening or closing quotes or parentheses (to avoid segmenting errors);

g. do not start with punctuation (to exclude improperly segmented fragments);

h. end in the kuten character, "。", the Japanese equivalent of a full stop or period (to include full sentences);

i. contain at least one predicate – verb or adjective (again, to exclude sentence fragments).

The intention of the above filters, obtained by empirical testing and evaluation, is to retain only well-formed, text-rich and relatively simple sentences.

Five subcorpora of different difficulty levels where then extracted from this collection of sentences, by selecting – for each subcorpus – only sentences with at least 10% of words belonging to the subcorpus difficulty level, and no words from a more difficult level. The size of the subcorpora is shown in Figure 2. Since both corpus and example collection were automated, relatively little manual labour was required to obtain a sizeable collection of examples.

Corpus	Sentences	Words	% jpWaCS-L2
jpWaCS	3,225,572	100,001,186	
jpWaCS-L2	859,416	13,395,667	100.00
jpWaCS-L2_0	351,935	5,536,969	40.95
jpWaCS-L2_1	34,777	403,470	4.05
jpWaCS-L2_2	96,161	1,172,911	11.19
jpWaCS-L2_3	26,894	264,979	3.13
jpWaCS-L2_4	9,830	79,473	1.14

Figure 2. JpWaC-L2 corpus and subcorpora contents

The corpora are available for Web concordancing at http://nl.ijs.si/jaslo/cqp/ through the search interface shown in Figure 3.

Users can choose to search the complete JpWaC-L2 corpus or only one subcorpus of the desired difficulty level. The "simple search" box can be used to search for any string (one or more words), while the "tabular search" section allows for combinations of searches of specific word forms, any form of a certain lemma, any word of a certain level, or any occurrence of a certain part of speech. The search result is a concordance where each line (sentence) is linked to its wider context within the original JpWaC

corpus, so that users can see the paragraph containing the sentence by clicking on the word. Concordances can be sorted on the left or right context, to investigate frequent collocational patterns and – in the case of verbs and adjectives – also flectional patterns. By selecting the option “Show: Word / Level / Lemma / Analysis” the user can choose to see the difficulty level, lemma and part-of-speech tag assigned by Chasen to each word in each concordance line, as seen in Figure 4.

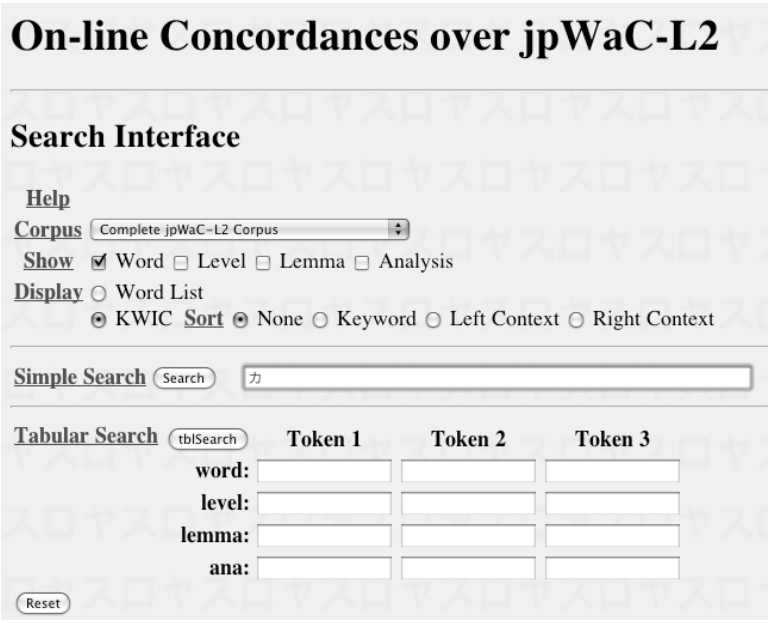


Figure 3. JpWaC-L2 corpus search interface

1	word	それ に は もう	慣れ た	。
	lemma	それ に は もう	慣れる た	。
	ana	N.Pron.g P.c.g P.bind Adv.g	V.free Aux	Sym.p
	level	4 4 4 4	3 4	-
2	word	早い 英語 に も だいぶ	慣れ た	ころ な のに 。
	lemma	早い 英語 に も だいぶ	慣れる た	ころ だ のに 。
	ana	Ai.free N.g P.c.g P.bind Adv.g	V.free Aux	N.bnd.Adv Aux P.Conj Sym.p
	level	4 4 4 4 3	3 4	4 4 3 -
3	word	この 生活 に も	慣れ た	と いう 事 。
	lemma	この 生活 に も	慣れる た	と いう 事 。
	ana	Adn N.Vs P.c.g P.bind	V.free Aux	P.c.r V.free N.bnd.g Sym.p
	level	4 3 4 4	3 4	4 4 3 -

Figure 4. JpWaC-L2 concordance for the verb form “nareta”

3.2. Evaluation of extracted examples

A sample of 10 lemmas for each level (4 to 0), for a total of 50 lemmas, was randomly extracted to evaluate the quantity and quality of examples in the corpus.

The average number of examples for these lemmas was 717 examples when searching through the whole JpWaC-L2 corpus, 497 examples in the level 0 subcorpus, 80 in the level 1 subcorpus, 278 in the level 2 subcorpus, 134 in the level 3 subcorpus, and 73 in the level 4 subcorpus, indicating that a sufficient amount of examples was found in each subcorpus.

Evaluating the grammaticality and acceptability of extracted sentences, it was found that less than 5% of the sentences were ungrammatical (containing garbled content or evident mistakes). A small percent of sentences were found to be assigned to the wrong lemma, due to Chasen's lemmatization error, and consequently sometimes also to the wrong difficulty level. The vast majority of the sentences, however, were well formed. The difficulty level assigned to the sentences, as measured according to the vocabulary contained, generally reflected their readability and comprehensibility. Shorter sentences were sometimes found not to be very informative without a wider context, while longer sentences, even if containing only basic vocabulary, sometimes contained challenging multi-word idiomatic expressions and syntactic structures. Although context for short sentences can be retrieved with a click, the criteria which define sentence length need further investigation.

4. Conclusion and further work

Two projects for the extraction of word usage examples from a parallel and a monolingual corpus were presented. In both cases, existing resources and automated processes were used to produce a collection of examples with relatively little manual labor. Plans for further work include a usability study, parallel corpus enlargement, and an enhancement of the selection procedure (applying criteria proposed by Mizuno *et al*. 2008 and Nishina and Yoshihashi 2007) and the measurement of example typicality, which has not yet been addressed by previous research on Japanese dictionary example selection, both in terms of vocabulary (collocations) and in terms of structure (morphological and syntactic patterns).

References

A. Dictionaries

BUNKACHÔ. (1971). *Gaikokujin no tame no kihongo yourei jiten – Dictionary of basic Japanese usage for foreigners*. Tokyo: Oogurashô insatsukyoku – Ministry of Finance Printing Bureau.

GROUP JAMASHII (1998). *Nihongo bunkei jiten*. Tokyo: Kurosio.

HIROSE, M. and SHOJI, K. (1994). *Effective Japanese Usage Guide*. Tokyo: Kodansha.

KAI, M. (2002). *Shôgaku shin kokugo jiten*. Tokyo: Mitsumura kyôiku tosho.

KITAHARA, Y. (2002). *Meikyô kokugo jiten*. Tokyo: Taishukan shoten.

MAKINO, S., NAKADA, S., OHSO, M. and JACOBSON, W.M. (1999). *Kodansha's Basic English-Japanese Dictionary*. Tokyo: Kodansha.

MATSUMURA, A. (2006). *Daijirin. Dai 3 han*. Tokyo: Sanseido.

MINATO, Y. (2008). *Challenge shôgaku kokugo jiten. Dai 4 han shin dezain han*. Tama: Benesse corporation.

NIHONGO NO KAI (1995). *Informative Japanese Dictionary*. Tokyo: Shinchosha.

SHARPE, P. (2006). *Kodansha's Communicative English-Japanese Dictionary*. Tokyo: Kodansha.

SHINMURA, I. (2008). *Kôjien. Dai 6 han*. Tokyo: Iwanami Shoten.

TAJIKA, J. (2009). *Sanseidô reikai shôgaku kokugojiten. Dai 4 han*. Tokyo: Sanseido.

TAKANO, T. (2004). *Gaikokujin no tame no tanoshii nihongo jiten – Japanese dictionary, learning language the fun way*. Tokyo: Sanseidô.

YAMADA, T. (2005). *Shin meikai kokugo jiten. Dai 6 han*. Tokyo: Sanseido.

B. Other references

ATKINS, S. and RUNDELL, M. (2008). *Oxford guide to practical* lexicography. Oxford: Oxford University Press.

ERJAVEC, T., HMELJAK SANGAWA, K. and SRDANOVIĆ ERJAVEC, I. (2006). jaSlo, A Japanese-Slovene Learners' Dictionary: Methods for Dictionary Enhancement. In E. Corino *Proceedings of the 12th EURALEX International Congress*. Alessandria: Edizioni dell'Orso: 611-616.

ERJAVEC, T., IGNAT, C., POULIQUEN, B. and STEINBERGER, R. (2005). Massive multi-lingual corpus compilation: Acquis Communautaire and totale. In *Proceedings of the 2nd Language & Technology Conference*, April 21-23, 2005, Poznan: Wydawnictwo Poznańskie: 32-36.

FOX, G. (1987). The case for examples. In J. Sinclair (ed.). *Looking up. An account of the Cobuild project in lexical computing*. London: Collins: 37-49.

Japan Foundation and Association of International Education Japan. (2004). *Japanese Language Proficiency Test. Test Content Specification*. Tokyo: Bonjinsha.

KAWAMURA, Y. and KANENIWA, K. (2006). Kokusai kyôdô henshû ni yoru nihongo gakushûsha no tame no tagengoban web jisho no kaihatsu (Development of a multilingual web-dictionary for learners of Japanese through international collaborative editing). In Nihongo Kyôiku Gakkai (ed.). *2006nendo nihongo kyôiku gakkai shunki taikai yokôshû (Proceedings of the 2006 Japanese language teaching association spring conference)*. Tokyo: Nihongo kyôiku gakkai: 61-66.

KILGARRIFF, A., RYCHLÝ, P., SMRŽ, P. and TUGWELL, D. (2004). The Sketch Engine. In G. Williams and S. Vessier (eds). *Proceedings of the Eleventh EURALEX International Congress*. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud: 105-116.

MAEKAWA, K. (2008). Compilation of the Balanced Corpus of Contemporary Written Japanese in the KOTONOHA Initiative. In *Proceedings of the Second International Symposium on Universal Communication – ISUC 2008*. Los Alamitos-Washington-Tokyo: IEEE: 169-172.

MATSUMOTO, Y., TAKAOKA, K. and ASAHARA, M. (2007). *Morphological analyzer chasen, version 2.4.0* {http://sourceforge.jp/projects/chasen-legacy/document/chasen-2.4.0-manual-j.pdf/ja/2/chasen-2.4.0-manual-j.pdf}.

MIZUNO, J., OOYAMA, H., KOBAYASHI, T., SAKATA, K., EVANS, N., TANIGUSHI, M. and MATSUMOTO, Y. (2008). Nihongo dokkai shien no tame no gogigoto no yôrei chûshutsu shisutemu no kôchiku. In *Proceedings of the Workshop on Natural Language Processing for Education – The 14th Annual Meeting of the Association for Natural Language Processing*. Tokyo: Gengo shori gakkai: 31-35.

NISHINA, K. AND YOSHIHASHI, K. (2007). Japanese composition support system displaying co-occurrences and example sentences. In S. Furui (ed.). *Proceedings of the Symposium on large-scale knowledge resources (LKR2007)*, Tokyo: Tokyo Institute of Technology: 119-122.

RYCHLÝ, P., HUSÁK, M., KILGARRIFF, A., RUNDELL, M. and MCADAM, K. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In E. Bernal and J. DeCesaris (eds). *Proceedings of the XIII EURALEX International Congress*. Barcelona: Institut Universitari de Lingüística Aplicada: 425-432.

SRDANOVIĆ ERJAVEC, I., ERJAVEC, T. and KILGARRIFF, A. (2008). A web corpus and word sketches for Japanese. *Journal of Natural Language Processing* 自然言語処理 15/2: 137-159.

TAMAMURA, F. (1984). *Goi no kenkyuu to kyouiku*, Tokio: Kokuritsu Kokugo Kenkyûjô.

Cultural values in a learner's dictionary: in search of a model

Olga Karpova[1], Mikhail Gorbunov[2]
Ivanovo State University

Abstract

The given paper presents the dictionary of a new explanatory-encyclopedic type addressed to guides and tourists. The main features of the dictionary and their implementation are described. Also the peculiarities of its electronic basis are specified. The given reference book is proposed as a model for building similar dictionaries.

Keywords: explanatory-encyclopedic, dictionary of a new type, Florence, electronic, guides and tourists, user profile.

1. Introduction

The modern lexicographic scene of the 21st century is characterized by a great variety of dictionary types. Such diversity can be explained by users' needs and demands. As the user profile has changed due to population migration in the world, the "native speaker" concept has also changed and now includes not only indigenous people but also a considerable number of immigrants (Karpova and Kartashkova 2009: 169-180).

Due to the changes in language situation and, consequently, the change of user profile, applied lexicography has extended its sphere and has turned into a science of compiling different types of reference works that once belonged neither to linguistic nor encyclopedic dictionaries. There also appeared another group of reference books – indices, calendars, and other resources and materials whose compilation demanded the use of lexicographic experience. Thus, the 21st century lexicography may be called "reference science". It combines both the theory of lexicography and the practice of compiling paper and electronic dictionaries of different types.

A lot of cultural notions demand explanation and other additional information (Karpova and Kartashkova 2007: 68-75). Considering modern users' inclination for receiving encyclopedic knowledge in one volume, authors of general and specialized dictionaries of modern English register in wordlists a significant number of toponyms and personal proper names from the most varied areas of knowledge.

[1] olga.m.karpova@gmail.com
[2] mvg85@ya.ru

The aim of many specialized dictionaries is to register and describe certain groups of words such as borrowings, proper names, terms. Dictionaries which register proper names are of special interest as they play an important role in cross-cultural studies. It is common knowledge that semantically proper names are viewed as a unity of linguistic and extralinguistic information.

In word-lists of such dictionaries users can find different thematic categories of toponyms, acquainting themselves with the author's creative works and the time he lived in. The culturological aspect in this type of dictionary is therefore essential (Karpova and Kartashkova 2009: 28-43).

Proper names are not only included in wordlists of glossaries, there are also dictionaries of characters and place names. This type of reference source led to the idea of combining the characteristics of different types of dictionaries (*e.g.* glossaries, encyclopedia, dictionaries of characters and place names, Wikipedia-like projects) in a new explanatory-encyclopedic dictionary addressed to a specific user group.

The lexical meanings of the entries are combined with encyclopedic definition and multimedia contents connected with the realities described. Such integrated lexicographic work would satisfy the requirements of modern users and the increased cultural requirements of our epoch.

2. A new type of dictionary

The dictionary project "Florence in the Works of European Writers and Artists. Encyclopedic Dictionary for Guides and Tourists"[3] is an example of such a combined dictionary (Karpova 2009: 20). It is addressed to a concrete user group of guides and tourists who want to obtain in-depth knowledge of Florence. Taking into consideration the possibilities and popularity of the Internet among modern users the dictionary is internet-based.

Taking this into consideration, the main characteristics of the dictionary are the following:

- open and dynamic character of dictionary information;
- the system of hyper references;
- the possibility to observe the information of the dictionary not as a formal stock of facts, but as a communicative fruitful system which makes the user of the dictionary a member of the whole process by giving him or her the possibility to extract knowledge from the suggested information.

[3] In cooperation with the Romualdo Del Bianco Foundation, international students' workshops "Florence in the Works of European Writers and Artists: Project of Encyclopedic Dictionary for Guides and Tourists" are organized every year since spring 2008.

This dictionary is structured according to theoretical rules accepted in applied lexicography (Bejoint 2000) and includes the following steps:

- Choosing a city as an object of cultural analysis;
- Specifying the sources of the dictionary (encyclopedias, dictionaries, biographies, guide books, etc.);
- Gathering data for the dictionary corpus;
- Designing the entry structure with special reference to its metalanguage;
- Creating hypertext links;
- Adding personal attitude to the entry described (the entry author describes his or her own impressions about the object defined in the entry, if he or she saw it personally);
- Creating a forum, where all the users and authors can propose their ideas about improving the dictionary.

The sources for this reference book are the following: the author's works and criticism of them, the author's biography from various printed and electronic encyclopedias and other reference sources, his notes and memoirs, the memoirs of his friends and relatives, the author's correspondence, etc.

The macrostructure of the dictionary consists of place names and biographical data of famous people from this region. Every entry word is provided with hyperlinks to other internet resources containing essential information about this person or object (museums, places, countries, cities, people and their works, books).

The dictionary microstructure consists of four sections: Biography, Creative Works, Florentine Influence, Learn more.

The entry includes the following information categories: graphic illustration (a photo, a picture, a portrait, etc.), chronological label (the date of birth of the author and/or date of creation of piece of art), encyclopedic definition; verbal illustrative examples (quotations, statements, sayings, etc).

Figure 1 illustrates the enumerated categories.

3. Characteristics of an electronic version of the dictionary

Each entry represents an unfolded hypertext containing references to dictionaries, encyclopedias and other reference resources dealing with certain places, writers and their works.

Multimedia features of the dictionary provide the user with additional possibilities (*e.g.* visual features such as a film, a picture or a video, and audio features such as a piece of music or song). The user can listen to a piece of music or watch a video presentation of the museum with the help of a built-in flash player plugin.

William Shakespeare
(1564–1616)

Biography

William Shakespeare was an English poet and playwright widely regarded as the greatest writer in the English language and the world's prominent dramatist. He is often called English national poet and the "Bard of Avon". Shakespeare was born and raised in Stratford-upon-Avon. At the age of 18, he married Anne Hathaway, who bore him three children: Susanna, and twins Hamnet and Judith. Between 1585 and 1592, he began a successful career in London as an actor, writer, and part owner of a playing company called the Lord Chamberlain's Men, later known as the King's Men. He appears to have retired to Stratford around 1613, where he died three years later.

Creative Works

His plays have been translated into every major living language and are performed more often than those of any other playwright. Shakespeare's surviving works consist of 36 plays, 154 sonnets, two long narrative poems, and several other poems. He produced most of his known works between 1590 and 1613. His early plays were mainly comedies and histories, genres he raised to the peak of sophistication and artistry by the end of the XVI century. He then wrote mainly tragedies until about 1608, including *Hamlet*, *King Lear*, and *Macbeth*, considered the finest works in the English language. In his last phase, he wrote tragicomedies, also known as romances, and collaborated with other playwrights.

Florentine Influence

Shakespeare's creative work was connected with ***Florence***.

The Oxford Companion to Shakespeare. R. A. Foakes, 2005 contains the following article on ***Florence***:

Florence, the capital of Tuscany, figures in *All's Well That Ends Well* (Florence is the setting of 3.5 and successive scenes). Florence is also mentioned in *The Taming of the Shrew* (1.1.14 and 4.2.91) and 'Florentines' (people from Florence) *in Much Ado About Nothing* (1.1.10) and *Othello* (1.1.19 and 3.1.39).

Who's Who in Shakespeare. A Dictionary of Characters and Proper names. F. G. Stokes, 1924 (2007) has the following entries:

Florence. Capital of Tuscany. For 'Duke of Florence': *All's Well*, i, 2. Mtd., *ib.* iii, 2; iv, 3 (2); v, 3 (2). 'Vincentio's son, brought up in Fl.' (*Tam. Sh.* i, 1); 'I have bills . . . by exchange from Fl.' (*ib.* iv, 2). 'Marcus Luccicos . . . in Fl.' (*Oth.* i, 3).

Florentine. (*a*) A native of Florence. 'The F.s and Senoys' (*All's Well*, i, 2}; 'a troop of F.s' (*ib.* iii, 6); mtd., *ib.* v, 3 (2); (Claudio) 'a young F.' (*M. Ado*, i.1); 'some F.' (*Tam. Sh.* i, 1); (Cassio) 'a F.' (*Oth.* i, 1); 'a F. [not Iago]' (*ib.* iii, 1).

(β) Duke of Florence. *All's Well*, i, 2; iv, 1, 3.

Learn more…

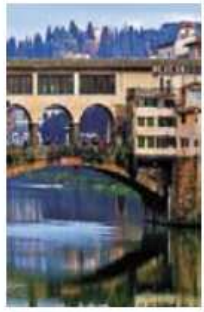


Figure 1. William Shakespeare, by Margarita Kulagina, Ivanovo State University

The electronic medium offers new opportunities (de Schryver 2003):

- simpler query interface;
- multiple layouts and new presentation formats;
- corpus-based data provision;
- new access possibilities;
- different multimedia devices (*e.g.* audio, video).

Such informational categories and multimedia features will provide the user with the necessary information about the given country and its culture, that is why the given reference book has some features of learner's dictionaries.

Such features give users possibilities in obtaining a great amount of information. A user-friendly interface will make the users' work fast and effective, giving pop-up prompts and additional data in the hints. Design of a microstructure implies user friendliness both in using and compiling the dictionary.

3.1. Customisable styles/colours

The styles system of the dictionary allows the appearance to be configured (*e.g.* colour, font, surrounding punctuation of each field in a dictionary). In the electronic dictionary, preconfigured "sets" of styles, or "views" of the data, may also be prepared for the end-user to choose between (*e.g.* "advanced" or "novice" views). Different fields may also be visible or hidden in different views. This provides for potentially highly flexible presentation of content based on end-user preference.

Special fonts and colour designations are used in the entries. The titles of the works are indicated in italics. The bold font is used in titles of the entry sections. This visually divides thematic blocks from each other and helps to speed up users' information search. All information on Florence is marked with red colour that draws users' attention and makes the reading process easier.

3.2. Customisable popup help

Each field – type of information – in the dictionary may optionally be linked to a popup "help screen" that is displayed to the user when clicking on that field. This behaviour is scriptable using the scripting language built into the dictionary allowing the popup help screen to exhibit some "intelligence", *e.g.* different information can be displayed depending on the type of field clicked on, as well as the value of a particular field.

3.3. User-friendly design: article preview

One of the primary design goals behind the development has been to produce a user-friendly tool: the software should be easy to learn and as intuitive as possible to use. One of the underlying principles for this goal is that lexicographers should not need to

have a high level of computer literacy in order to perform the day-to-day tasks of compiling a dictionary. The level of abstraction presented to the lexicographer should be that of a dictionary article, and not that of a database. The generic input/output architecture supports this idea by 'hiding', wherever possible, the technical details of how the dictionary is stored. A WYSIWYG editor is used for making the editing process simpler and quicker.

3.4. Customisable searching.

The ability to configure a search script using the built-in scripting language allows the search behaviour to be potentially customised for language-specific search functionality, if desired.

3.5. Full dictionary searches

A text search function allows the entire dictionary to be searched quickly for a particular piece of text. This includes options such as case-sensitivity, whole-word/partial-word matching, and support for regular expressions.

3.6. Extendible input/output architecture

The input/output (I/O) architecture of the dictionary is designed to support multiple types of data storage mechanisms. The primary data storage type for networked multi-user use is a relational database system, such as MySQL or Microsoft SQL. This is implemented internally using Object Database Connectivity (ODBC) and Structured Query Language (SQL). Some of the export – output – formats supported in the dictionary are static HTML – with or without style sheets – and Rich Text Format (RTF) for producing print output in word processors such as Microsoft Word, OpenOffice and Corel WordPerfect. The I/O architecture has been genericised, which allows additional interfaces for different types of data sources to be developed in the form of add-ons or plugins. This allows for the possibility of custom importers to be created in cases where there may be existing dictionary data developed in another system. Add-on modules may also be created in order to support other output formats.

3.7. Localisation and dynamic metalanguage customization

The entire interface of the dictionary can be translated to, and made available in, any language. The default language may also be changed. Thus, for example, the interface of a bilingual German-English dictionary could be available in both German and English, with German being the default. Instant "real-time" switching of the language allows the interface language to be changed by the end-user without even re-opening the page. Taking localisation further, the metalanguage of the dictionary content may also be localised. The end-user may opt to see labels or words in cross-references, in the language of their choice.

The data in the dictionary bank is assembled by a coordinated international group of students and academics at numerous higher schools. Still there are some crucial issues to be solved:

- whether there should be a corpus for such a dictionary; if yes, then
- what type of corpus it should be,
- how it is going to be collected and what tools are needed.

The answers to these questions will form the theoretical base for the new type of reference sources.

This dictionary has a unique feature which makes it possible to reflect in the dictionary personal impressions and associations provided by the entries' authors (participants of student's workshops in Florence). It contributes to cross-cultural learning experiences of the users and facilitates the educational process based on the study of foreign literature, arts and music.

4. Conclusion

The new type of reference resource described here (see also Karpova 2009) is characterized by the following innovative features (*cf*. Karpova 2008: 272-273):

- electronic corpora and Internet reference facilities;
- technical innovations and challenging design;
- integration of several lexicographic forms in one volume;
- new approaches to description of key-words;
- changing of users' profile and their new needs and demands.

It can serve as a model for creating analogous reference sources for any city and country regardless of the language, religion or culture.

References

BEJOINT, H. (2000). *Modern Lexicography: An Introduction*, Oxford U.P.

CARR. M. (1997). Internet Dictionaries and Lexicography. *International Journal of lexicography*, 10(3): 209-221.

DE SCHRYVER, G.-M. (2003). Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography*, 16(2): 143-199.

KARPOVA, O. (2004). *Author's Lexicography with Special Reference to Shakespeare Dictionaries. Historical Dictionaries and Historical Dictionary Research. Lexicographica. Series maior*. Band 123: 31-38.

KARPOVA, O. (2005). Russian Lexicography. In K. Brown (ed.). *Oxford Encyclopedia of Language &. Linguistics*. Vol. 10. Oxford: Oxford University Press: 704-715.

KARPOVA, O. AND KARTASHKOVA, F. (ed.) (2007). *Essays on Lexicon, Lexicography, Terminography in Russian, American and Other Cultures*. Cambridge Scholars Publishing.

KARPOVA, O. (2008). Dictionaries of Shakespeare Proverbs and Quotations. In R. Soares and O. Lauhakangas (eds). *Imagery of Proverbs: The Great Chain of Being as the Background of Personificatory and Depersonificatory Metaphors in Proverbs and Elsewhere.* Adas ICP07 Proceedings. Tavira: Tipografia Tavirense: 271-277.

KARPOVA, O. (2009). *Florence in the Works of European Writers and Artists. Encyclopedic Dictionary for Guides and Tourists*. Ivanovo State University.

KARPOVA, O. and KARTASHKOVA, F. (eds) (2009). *Lexicography and Terminology: A Worldwide Outlook.* Cambridge Scholars Publishing.

OII, V. (1998). *Computer Corpus Lexicography*. Edinburgh: Edinburgh University Press.

On the combination of automated information and lexicographically interpreted information in two German online dictionaries

Annette Klosa[1]
Institut für Deutsche Sprache, Mannheim

Abstract

This paper discusses the advantages and disadvantages of the combination of automated information and lexicographically interpreted information in online dictionaries, namely *elexiko*, a hypertext dictionary and lexical data information system of contemporary German (http://www.owid.de/elexiko_/index.html), and DWDS, a digital dictionary of 20th century German (http://www.dwds.de). Examples of automatically derived information (*e.g.* automatically extracted citations from the underlying corpus, lists on paradigmatic relations) and lexicographically compiled information (*e.g.* information on paradigmatic partners) are provided and evaluated, reflecting on the need to develop guidelines as to how computerised information and lexicographically interpreted information may be combined profitably in online reference works.

Keywords: online dictionary, German, automated information, lexicographically interpreted information.

1. Introduction

The availability of large electronic corpora has changed the work of lexicographers in more ways than one. By applying corpus-driven and corpus-based approaches (*cf.* Tognini-Bonelli 2001), thus exploiting large electronic text corpora, modern monolingual and bilingual dictionaries are developed. The ways in which this has affected the process of compiling a dictionary have been described for specific dictionary projects (*cf.* Baugh, Harley and Jellis 1996; Sinclair 1987) and reflected on in a broader approach in various publications (*cf.* Klosa 2007; Teubert 1999; Euralex Bibliography of Lexicography: http://euralex.pbwiki.com/Corpus+Lexicography, and OBELEX: http://hypermedia.ids-mannheim.de/pls/lexpublic/bib_en.ansicht; keyword=corpus-based lexicography).

But data from large corpora is not only interpreted by lexicographers: it is also the basis for computer linguistic tools and their applications such as collocation analysis. Computational linguists have developed procedures often based on statistical methods

[1] klosa@ids-mannheim.de

and designed to calculate frequency and significance, or for part-of-speech tagging. They have contributed, for example, to the development of lexical-semantic resources (*e.g.* ontologies, *cf.* Mönnich and Kühnberger 2008), or automatic sense disambiguation (*cf.* Agirre and Edmonds 2006). Computational lexicographers, in particular, have been concerned with questions of how to build a lexicon (*cf.* Boguraev 1993).

Within this context, the idea of automating and, thus, possibly accelerating the compilation of dictionaries has emerged. In the past, working on a dictionary was an exclusively human task. Nowadays, it is a combination of applying computer and corpus tools together with the lexicographer's linguistic competence. In printed dictionaries, this mainly leads to an improvement in the quality of lexicographic information but not necessarily to new types of lexicographic information. In electronic dictionaries, this can (and maybe even should) be different. Electronic dictionaries and online dictionaries, in particular, are not subjected to limitations of space. As well as the classic inventory of grammatical, morphological, orthographic, semantic, and pragmatic information, electronic dictionaries are able to offer more detailed information and new types of linguistic detail. They may also present data in non–traditional ways, for example in graphs or by using other media, such as video or audio files.

In this situation, lexicographers have to assess what computational linguistics can offer. They have to decide which information in the dictionary must still be manually compiled (*e.g.* the paraphrase of the headword) and which information might be automatically extracted (*e.g.* information on part of speech or inflection). The advantages and disadvantages of the combination of automated information and lexicographically interpreted information in online dictionaries are discussed here by looking at two German dictionaries: *elexiko*, a hypertext dictionary and lexical data information system of contemporary German (http://www.owid.de/elexiko_/index.html), and DWDS, a digital dictionary of 20th century German (http://www.dwds.de).

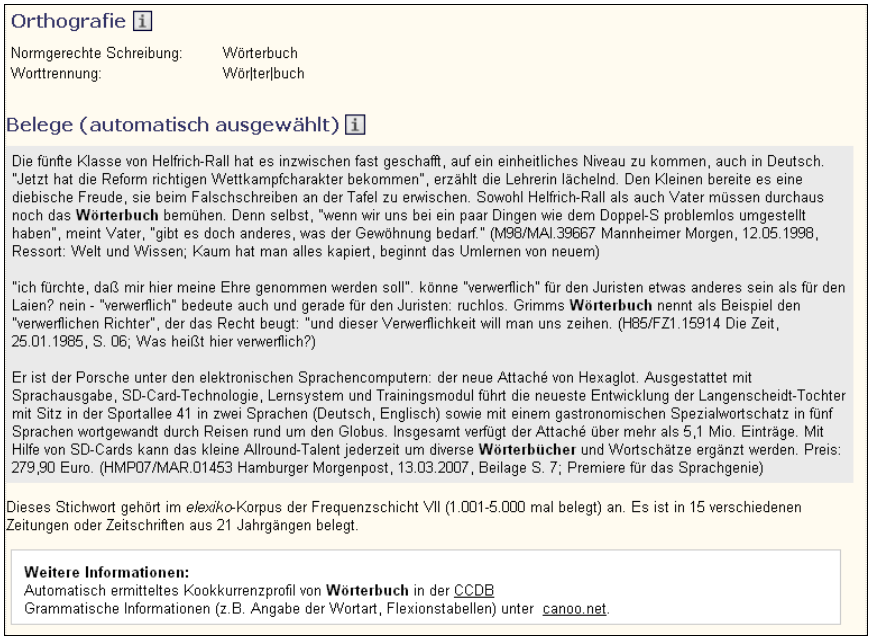
2. *elexiko*

elexiko is a lexicological-lexicographic project based at the Institute for the German Language (IDS) at Mannheim (*cf.* Haß 2005; Klosa, *et al.* 2006; Storjohann 2005b). The aim of this project is to compile a reference work, specifically designed for online publication, that explains and documents contemporary German. The primary and exclusive basis for lexicographic interpretation is an extensive German corpus. Filling *elexiko* in modules is (besides the corpus-based approach) one of the two main lexicographic methods for the dictionary. *elexiko* is compiled not in alphabetical order but by analysing the semantic, syntactic, or morphological features of the lexicon systematically in batches. Thus, a complete word class, an entire word family, or a semantic field can be described systematically and separately. Furthermore, modules are also defined according to levels of frequency and distribution of lexemes in the

elexiko corpus. Right now, complex and comprehensive information on a module called “Dictionary on Public Discourse” is being compiled. It contains approximately 2,800 entries selected mainly according to their (high) frequency in the *elexiko* corpus.

Along with publishing the list of headwords (taken exclusively from the *elexiko* corpus) on the Internet in 2003, the *elexiko* dictionary was filled with sense-independent information for each headword generated automatically from the underlying corpus. This concerns 300,000 single-word entries comprising details on spelling, spelling variation, and syllabication. The orthographic information in particular was checked manually, because mistakes here are usually not tolerated by users. Since then, the project has been working on enriching as many entries as possible with further information generated automatically, *e.g.* automatically chosen citations (see Figure 1).

Choosing citations from the corpus is not carried out strictly according to rules of statistical concurrence, but by applying certain criteria, which help to improve the quality of the citations. For example, they have to be found in at least three different sources and come from at least three different years. Users may find these quotations helpful when they look up the meaning of a word. In addition to these text clippings, information on the coverage of the headword in the *elexiko* corpus is given. By showing the number of sources and years in which the headword occurs in the corpus, the user may get an idea of the distribution of the word.



Orthografie

Normgerechte Schreibung: Wörterbuch
Worttrennung: Wör|ter|buch

Belege (automatisch ausgewählt)

Die fünfte Klasse von Helfrich-Rall hat es inzwischen fast geschafft, auf ein einheitliches Niveau zu kommen, auch in Deutsch. "Jetzt hat die Reform richtigen Wettkampfcharakter bekommen", erzählt die Lehrerin lächelnd. Den Kleinen bereite es eine diebische Freude, sie beim Falschschreiben an der Tafel zu erwischen. Sowohl Helfrich-Rall als auch Vater müssen durchaus noch das **Wörterbuch** bemühen. Denn selbst, "wenn wir uns bei ein paar Dingen wie dem Doppel-S problemlos umgestellt haben", meint Vater, "gibt es doch anderes, was der Gewöhnung bedarf." (M98/MAI.39667 Mannheimer Morgen, 12.05.1998, Ressort: Welt und Wissen; Kaum hat man alles kapiert, beginnt das Umlernen von neuem)

"ich fürchte, daß mir hier meine Ehre genommen werden soll". könne "verwerflich" für den Juristen etwas anderes sein als für den Laien? nein - "verwerflich" bedeute auch und gerade für den Juristen: ruchlos. Grimms **Wörterbuch** nennt als Beispiel den "verwerflichen Richter", der das Recht beugt: "und dieser Verwerflichkeit will man uns zeihen. (H85/FZ1.15914 Die Zeit, 25.01.1985, S. 06; Was heißt hier verwerflich?)

Er ist der Porsche unter den elektronischen Sprachencomputern: der neue Attaché von Hexaglot. Ausgestattet mit Sprachausgabe, SD-Card-Technologie, Lernsystem und Trainingsmodul führt die neueste Entwicklung der Langenscheidt-Tochter mit Sitz in der Sportallee 41 in zwei Sprachen (Deutsch, Englisch) sowie mit einem gastronomischen Spezialwortschatz in fünf Sprachen wortgewandt durch Reisen rund um den Globus. Insgesamt verfügt der Attaché über mehr als 5,1 Mio. Einträge. Mit Hilfe von SD-Cards kann das kleine Allround-Talent jederzeit um diverse **Wörterbücher** und Wortschätze ergänzt werden. Preis: 279,90 Euro. (HMP07/MAR.01453 Hamburger Morgenpost, 13.03.2007, Beilage S. 7; Premiere für das Sprachgenie)

Dieses Stichwort gehört im *elexiko*-Korpus der Frequenzschicht VII (1.001-5.000 mal belegt) an. Es ist in 15 verschiedenen Zeitungen oder Zeitschriften aus 21 Jahrgängen belegt.

Weitere Informationen:
Automatisch ermitteltes Kookkurrenzprofil von **Wörterbuch** in der CCDB
Grammatische Informationen (z.B. Angabe der Wortart, Flexionstabellen) unter canoo.net.

Figure 1. Automated information in elexiko on the headword “Wörterbuch”

Additionally in *elexiko*, there are hyperlinks to other online sources, where users may look up automatically compiled information on collocations (hyperlink to "Kookkurrenzdatenbank CCDB" developed at the IDS) and on grammar (hyperlink to canoo.net, where information on flexion and word formation is given). A direct link from dictionary entries to the underlying corpus has not yet been implemented, because many of the corpus texts are not openly accessible due to copyright. In the near future, *elexiko* will offer automatically compiled information on word formation with the headword. Words stemming from one headword will be given as hyperlinks, thus joining entries with lexicographically and automatically compiled information.

In *elexiko*, automated information is employed carefully; as much of this information as possible is checked manually in order to improve the quality. This has the negative effect of slowing down the process of publication and increasing the cost.

3. DWDS – Digital Dictionary of contemporary German

DWDS, a digital dictionary of contemporary German published at the Berlin-Brandenburg Academy of Science since 2004, was planned differently from the start (*cf.* Klein and Geyken 2000, 2001; Geyken 2005). This project aims at creating a "digital lexical system", that is easy to expand or correct and may be used for many different academic or non-academic purposes. DWDS combines a digitalised print dictionary with a word profile giving automated information on collocations, citations from an extensive corpus on German between 1900 and 1990, and a thesaurus. In the beta version of DWDS, which is shown here, on the first screen after looking up a word, all this information is combined (see Figure 2).

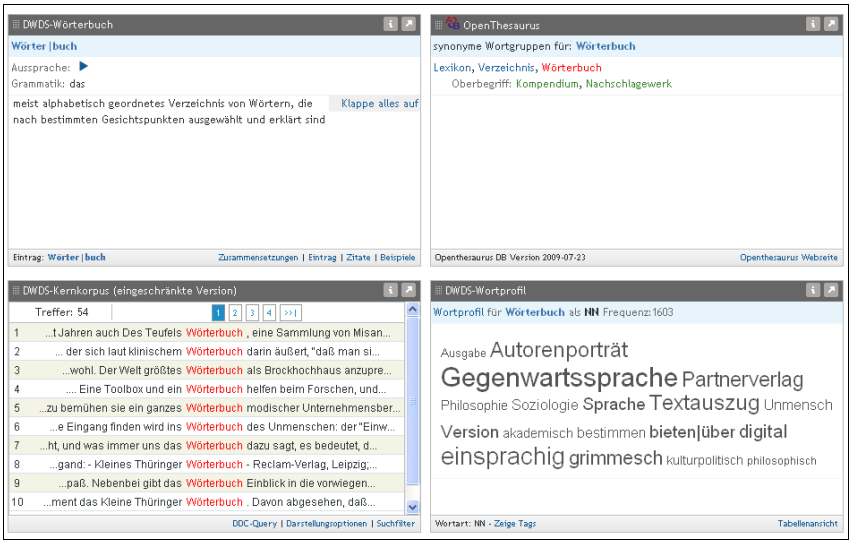


Figure 2. Information in DWDS on the headword "Wörterbuch"

The "Dictionary of Contemporary German" (WDG) was digitalised for DWDS, but was published in the 1960s and 70s in the German Democratic Republic and has been written on the basis of a paper archive of citations. For the online version, the possibility of presenting only part of the information in the entry has been implemented, and information on pronunciation will soon be added.

The thesaurus incorporated is not developed by DWDS, but OpenThesaurus maintains its own domain, where anybody may contribute to the dictionary. Its information is built into the DWDS site, but not hyperlinked with other information. Synonyms given, for example, are not hyperlinked to entries in the WDG dictionary.

The citations quoted come from the DWDS corpora and are chosen automatically. Usually 10 KWICs are given, but full contexts and more KWICs may be opened. Although the DWDS corpora were planned carefully, the quality of the citations given is not always convincing, as with any automatic selection. But even more important is that the DWDS corpora were not the basis for the WDG dictionary.

The word profile gives words and phrases collocating with the headword in the DWDS corpora in a word cloud. The word cloud shows the most frequent words co-occurring with the headword, but in many cases those words are not part of the WDG dictionary entry itself and vice versa. This is of course the case because the DWDS corpora are not the basis for the WDG dictionary. Here, automated information from one source and lexicographically written information from another source do not really harmonise, but at least they complement each other.

4. Conclusion

Two online reference works for German approach the matter of incorporating automated and lexicographically compiled information on words completely differently. While DWDS has compiled a lot of information on German from different sources in quite a short time and presents it in one user interface without tagging each kind of information, *elexiko* has less information, but hyperlinks to other applications. Automatically compiled information is checked lexicographically in *elexiko* as much as possible, slowing down the process of publication. Automated information is also labelled as such. In addition to this, new corpus-based, complex and comprehensive information on very frequent entries is being compiled in *elexiko*.

When contrasting information on paradigmatic relations in DWDS and *elexiko*, the huge difference in quality and quantity between automated information and lexicographically compiled information becomes apparent. The OpenThesaurus in DWDS gives the following synonyms (single words and multi-word units) for the headword *Aids* (http://beta.dwds.de/?qu=Aids&view=1): *Acquired Immune Deficiency Syndrome*, *AIDS*, *erworbenes Immunschwäche-Syndrom* (*i.e.* 'acquired immune deficiency syndrome'), and it records *Infektionskrankheit* (*i.e.* 'infectious disease') as a hypernym. In its word profile, DWDS names the collocates *Armut* (*i.e.* 'poverty'), *Malaria* (*i.e.* 'malaria'), *Tuberkulose* (*i.e.* 'tuberculosis') and *sterben an* (*i.e.* 'to die

of'). Three of these collocates would probably be classified as paradigmatic partners in *elexiko*, the verbal phrase *an Aids sterben* (*i.e.* 'to die of Aids') would appear as typical usage.

In the *elexiko* entry for *Aids* (http://www.owid.de/pls/db/p4_anzeige.lesart?v_id=302141&v_lesart=Krankheit) synonyms given are *Immunschwäche* (*i.e.* 'immune deficiency') and *Immunschwächekrankheit* (*i.e.* 'immune deficiency disease'), hypernyms are *Epidemie* (*i.e.* 'epidemic'), *Erkrankung* (*i.e.* 'disease'), *Krankheit* (*i.e.* 'illness'), *Infektionskrankheit* (*i.e.* 'infectious disease') and *Seuche* (*i.e.* 'epidemic'). There are also three thematically defined groups of incompatible (*i.e.* cohyponym) partner words: *Armut* (*i.e.* 'poverty') and *Hunger* (*i.e.* 'hunger'), *Geschlechtskrankheit* (*i.e.* 'veneral disease') and *HIV-Infektion* (*i.e.* 'HIV infection'), *Alkohol* (*i.e.* 'alcohol') and *Droge* (*i.e.* 'drug'). Each paradigmatic partner is accompanied by a citation illustrating the relation (*cf.* Storjohann 2005a). In addition, information on each type of paradigmatic relation can be opened.

It is not for lexicographers to decide which way is to be preferred, but for the users. We do not yet know whether users would like to be able to rate the reliability of lexical information in online dictionaries. We do not even know how users will respond to automated information in a dictionary in general. Will users appreciate the comprehensive description of paradigmatic relations in *elexiko* or will automated information as in DWDS satisfy them in specific instances of dictionary use? Lexicographers can only assess the possibilities computational linguistics offers, and test new ways of enriching electronic dictionaries with lexicographically compiled and automated information, linking them in a fruitful way. Only extensive usage research, as intended for the *elexiko* project, will help to find answers to these questions.

References

AGIRRE, E. and EDMONDS, P. (eds). (2006). *Word sense disambiguation: Algorithms and applications*. Dordrecht: Springer.

BAUGH, S., HARLEY, A. and JELLIS, S. (1996). The Role of Corpora in Compiling the Cambridge International Dictionary of English. *International Journal of Corpus Linguistics* 1/1: 39-59.

BOGURAEV, B.K. (1993). The contribution of computational lexicography. In M. Bates and R.M. Weischedel (eds). *Challenges in Natural Language Processing*. Cambridge: Cambridge University Press: 99-134.

DWDS – Digital Dictionary of the German Language of the 20th Century: http://www.dwds.de/ (02.11.2009).

elexiko: http://www.owid.de/elexiko_/index.html (02.11.2009).

Euralex Bibliography of Lexicography: http://euralex.pbwiki.com/ (02.11.2009).

GEYKEN, A. (2005). Das Wortinformationssystem des Digitalen Wörterbuchs der deutschen Sprache des 20. Jahrhunderts (DWDS). *BBAW Circular*, 32: 40.

HAß, U. (2005). *Grundfragen der elektronischen Lexikographie. elexiko – Das Online-Informationssystem zum deutschen Wortschatz*. Berlin/New York: de Gruyter.

KLEIN, W. and GEYKEN, A. (2000). Projekt «Digitales Wörterbuch der deutschen Sprache des 20. Jh.». *Jahrbuch der BBAW*, 1999: 277-289.

KLEIN, W. and GEYKEN, A. (2001). Projekt «Digitales Wörterbuch der deutschen Sprache des 20. Jh. ». *Jahrbuch der BBAW*, 2000: 263-270.

KLOSA, A. (2007). Korpusgestützte Lexikographie: besser, schneller, umfangreicher? In W. Kallmeyer and G. Zifonun (eds). *Sprachkorpora. Datenmengen und Erkenntnisfortschritt*. Berlin/New York: de Gruyter: 105-122.

KLOSA, A., SCHNÖRCH, U. and STORJOHANN, P. (2006). ELEXIKO – A lexical and lexicological, corpus-based hypertext information system at the Institut für Deusche Sprache, Mannheim. In C. Marello *et al.* (eds). *Proceedings of the 12th EURALEX International Congress* (Atti del XII Congresso Internazionale di Lessicografia), EURALEX 2006, Turin, Italy, September 6th-9th, 2006. Vol. 1. Turin: Edizioni dell'Orso Alessandria: 425-430.

MÖNNICH, U. and KÜHNBERGER, K.-U. (eds). (2008). *Foundations of Ontologies in Text Technology, Part II: Applications*. (*Zeitschrift für Computerlinguistik und Sprachtechnologie*, 23/1).

OBELEX – Online Bibliography of Electronic Lexicography: http://hypermedia.ids-mannheim.de/pls/lexpublic/bib_en.ansicht (02.11.2009).

SINCLAIR, J. (1987). *Looking Up. An Account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary*. London: Harper Collins.

STORJOHANN, P. (2005a). Corpus-driven *vs*. corpus-based approach to the study of relational patterns. In *Proceedings of the Corpus Linguistics Conference 2005 in Birmingham. Vol. 1, no. 1*. (http://www.corpus.bham.ac.uk/PCLC/).

STORJOHANN, P. (2005b). *elexiko* – A Corpus-Based Monolingual German Dictionary. *Hermes, Journal of Linguistics*, 34: 55-83.

TEUBERT, W. (1999). Korpuslinguistik und Lexikographie. *Deutsche Sprache*, 4: 292-313.

TOGNINI-BONELLI, E. (2001). *Corpus Linguistics at Work*. Amsterdam and Philadelphia: Benjamins.

Making a dictionary without words: lemmatization problems in a sign language dictionary

Jette Hedegaard Kristoffersen[1], Thomas Troelsgård[1]
Centre for Sign Language, University College Capital, Copenhagen

Abstract

This paper addresses some of the particular problems connected with lemma representation and lemmatization in a sign language dictionary. The paper is mainly based on the authors' work experience from the Danish Sign Language Dictionary project. In a sign language dictionary sign representation constitutes a problem, as there is – at least for Danish Sign Language – no conventional notation used by native signers and the various other sign user groups. We look into the different possibilities of representing signs and present the solution that we chose for the Danish Sign Language Dictionary.

Defining the criteria for lemmatization is another area where sign language dictionaries differ from written language dictionaries. The criteria should obviously include the manual expression of the signs, but a sign's manual expression has features from several categories (*e.g.* handshape, place of articulation and movement). Also non-manual elements such as mouth movement could be taken into consideration when defining the lemmatization criteria. As we defined the lemmatization criteria for the Danish Sign Language Dictionary we aimed for a solution that would result in relatively few homonyms, but that at the same time would not lead to very large polysemous entries. We also tried to define the criteria so that the resulting entries would reflect the lexicon of Danish Sign Language rather than resembling a Danish dictionary.

Keywords: Danish Sign Language, sign language lexicography, sign language dictionary, lemma representation, lemmatization.

1. Introduction

The Danish Sign Language Dictionary project was a five year project that was concluded in 2008 with the publication of a dictionary with approximately 2,000 sign entries. The entries cover about 3,000 meanings with 6,000 Danish equivalents and 3,500 usage examples. The dictionary is freely accessible online at www.tegnsprog.dk.

The main element of sign language is obviously the manual expression – the actual signs, but sign languages also comprise a series of non-manual features such as facial expression, mouth movement, eye-gaze, eye-blink and movement of the upper body. Danish Sign Language (DTS) is in many ways influenced by Danish, but it is a full, independent language, with its own lexicon, morphology, syntax, etc.

[1] {jehk,ttro}@ucc.dk

2. Sign language notation

For a spoken and written language like English or Danish there are norms for spelling, and although some word-forms have competing variants, most words have a spelling and a base form, which the majority of the language users agree on.

For DTS, and probably for many other sign languages, this is not the case. There is no standard notation system adopted by the native signers. This causes trouble for the lexicographer, who of course would like to work with a fixed unique representation of the lemma that can easily be read.

2.1. Existing notation systems for signed languages

There exist several formal notation systems for sign languages. In the following we will show three of these, using the DTS sign for 'temperature' as an example (see Figure 1).

Figure 1. DTS sign for 'temperature'

The first system used for ordering a dictionary was developed by William Stokoe in the mid 20th century and used in the "Dictionary of American Sign Language" from 1965. This system has three sets of symbols describing location, handshape and movement. For an example of the Stokoe system, see Figure 2 which represents the entry for 'temperature' (same sign as in Figure 1) in the 1965 American Sign Language (ASL) dictionary (Stokoe *et al*. 1965).

G^ ♀ G< N/X
(imit.; tab may be B^) N *temperature.*

Figure 2. Entry for 'temperature' *in the 1965 ASL dictionary*

A newer system is The Hamburg Notation System (HamNoSys) developed at the Hamburg University. This system is based on the same principles as the Stokoe system, but it allows for more details and is more flexible. For a description of the

system, see *e.g.* Prillwitz *et al.* (1989). Figure 3 shows the example sign for ‘temperature’ written in HamNoSys.

Figure 3. The DTS sign for ‘temperature’ *written in HamNoSys*

This system is used not only at Hamburg University, but also among sign language learners, teachers and researchers in many places. It has however never been adapted either by the deaf community, or by other users of sign language as an everyday tool for reading and writing.

Another approach is used by the SignWriting system developed by Valerie Sutton in the 1970s, originally based on a system for dance notation. SignWriting renders signs as stylized pictogram-like drawings. Figure 4 shows the example sign for ‘temperature’ in SignWriting. For more information on SignWriting, see www.signwriting.org.

Figure 4. The DTS sign for ‘temperature’ *written in SignWriting*

With some exercise you can read and write SignWriting. However, the system has one great disadvantage in a digital context – in opposition to the Stokoe and Hamburg systems, it is not linear. A digital representation of SignWriting is a series of graphic elements that cannot be searched and ordered according to the individual manual elements that constitute the signs.

2.2. Other ways of rendering signs

Another way of representing signs is through a drawing or photo, potentially with arrows showing the sign movement (*cf.* Figure 1), or through several drawings and photos, showing the different sequences of the sign. Other possibilities of rendering signs are video recordings, textual descriptions of the sign production, and glosses – equivalents from written language, usually written in upper case.

2.3. Sign representation in the Danish Sign Language Dictionary

Among DTS learners and teachers there is some use of SignWriting, but the most common way to write signs is glosses. These glosses are not standardised, but made up along the way, based on meaning of the sign in context. None of the existing sign notation systems can easily be read and written by all of the very diverse user groups of the Danish Sign Language Dictionary. As a result, we chose to present the headwords as videos, that is, to use a direct rendering of the signs. In comparison with

the existing notation systems this is the most accurate and informative rendering of the signs and it demands no particular qualifications of the users.

The video solution works well in the sign entry head, but in the search result list, where several signs are shown together, and in situations where sign references are surrounded by text, *e.g.* in example sentences, cross-references, etc., it would obviously be quite confusing to use moving video clips. Therefore, we render signs as photographs and/or glosses in the entry body. In the entry head, on the other hand, both a photograph, a gloss, and a video clip is used for rendering the lemma (*cf.* Figure 5).

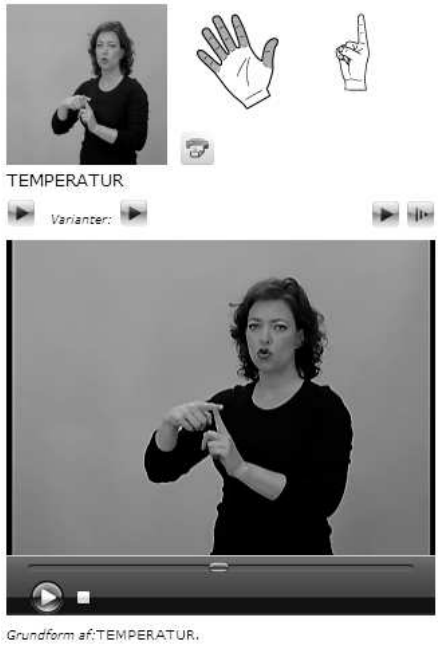


Figure 5. Entry for DTS sign for 'temperature' *in the Danish Sign Language Dictionary*

3. Searching and ordering a sign language dictionary

So far, we have discussed how to present the headword to the user. Another problem is connected to the search function of the dictionary. In the Danish Sign Language Dictionary it should be possible to look up the signs knowing only their manual form. It would also be expected that the signs in a search result list could be ordered according to their manual expression – that is "alphabetically", not by words and letters, but by the signs themselves. We therefore had to add information about the basic phonological features of each sign and its variants in our dictionary database.

These data are used by the searching and ordering functions in the dictionary, but are not shown in the actual entries.

4. Phonological features as lemmatization criteria

Lemmatization in the Danish Sign Language Dictionary is based partly on phonology, partly on semantics. A sign is traditionally described as a simultaneous unit of items from four different parameters: shape of the hand, orientation of the hand, location of the hand and movement of the hand. This way of analyzing signs was introduced by several researchers in the late 1970s (*e.g.* Klima and Bellugi 1979).

Within the framework of auto-segmental phonology Liddell and Johnson (1987) developed a model for phonological description of signs based on the notion that signs have a segmental structure. Figure 6 shows how each sign consists of at least one start position described by four parameters, one movement and one end position described by four parameters.

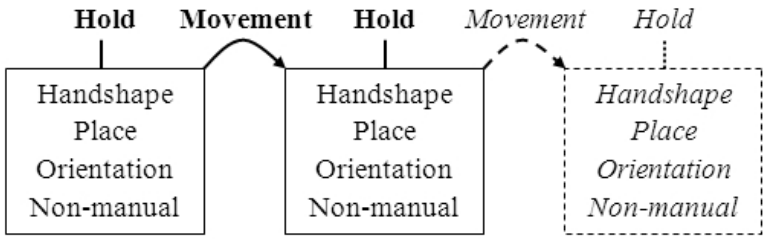


Figure 6. A segmental sign description model

A sign can have more than one movement and therefore additional hand shapes, orientations, places of articulation and non-manual features as indicated by the right-hand part of Figure 6. Furthermore, a sign can be articulated with two hands, which results in a two tiers in the description. All these features are taken into consideration when we do the lemmatization part based on the phonological features. Liddell and Johnson (1987) also included non-manual features such as facial expressions in the Hold-part.

Only a few signs in DTS have facial expression or eye gaze as a lexicalized part. Most facial expressions, eye gaze and body movements are suprasegmental units on phrase level and parts of the grammar. Every sign, however, has movement of the mouth as a lexical property, and the question for us has been whether to treat mouth movements as part of the phonology – like the manual part of the signs – or to treat mouth movements as a part of the semantic description.

4.1. Mouth movements

Some signs have different mouth patterns due to the fact that DTS and Danish have different ways to separate concepts in real life. The manual part of the sign HUS ‘house’ has many possible mouthings – that is, silent imitations of the visible

articulation of a Danish equivalent: /hus/ (’house’), /ejendom/ (‘home’), /villa/ (‘villa’), /hal/ (‘hall’), /hytte/ (‘cottage’) and /klinik/ (‘clinic’). Mouth movements are not solely silent imitations of spoken Danish. DTS also has oral components not related to the surrounding spoken language, that is mouth movements that do not reflect Danish words, but are particular to DTS (Kristoffersen and Niemäla 2008).

The sign MASSER-AF, meaning ‘lots of’, ‘many’ and ‘how many’ can be accompanied by four of the five Danish equivalents given in the Danish Sign Language Dictionary and by the three mouth gestures shown in Figures 7, 8 and 9.

Figure 7. Mouth gesture [i]

Figure 8. Mouth gesture <Blowing>

Figure 9. Mouth gesture <vibrating lips>

If mouth patterns were described as a part of the phonological description, the sign HUS would be split into six different lemmas and the sign MASSER-AF would be

split into seven different lemmas. If, however, mouth patterns were described at the meaning level, HUS and MASSER-AF would be treated as two signs.

To split the signs into six and seven different sign entries seemed contra-intuitive both to the 25 naïve native signers (of a total population of 4,000 native signers) in our group of consultants, and to the two native signers and trained linguists in our lexicographic staff.

We decided to locate the description of mouth patterns at the meaning level, thus obtaining the possibility of describing these without being forced to add new entries to the dictionary in cases like HUS and MASSER-AF. Consequently, an entry is phonologically established solely by the headword's manual expression, that is with no reference to oral (or other non-manual) features.

Some signs cannot be accompanied by all their Danish equivalents as mouth movements. For example, the manual part of the sign HUS 'house' may be translated into eight Danish synonyms, but the sign can only be accompanied by the mouthing of seven of these. The Danish Sign Language Dictionary aims to show which Danish mouthings the individual signs can be accompanied by. Every Danish equivalent is marked in case the imitation of the equivalent is an acceptable mouthing accompanying the sign. Additionally, oral components that are not mouthings of Danish equivalents are shown in the entry.

4.2. Variant or synonym?

As already mentioned, there is no standard notation system adopted by the native signers, and being a language with no written representation results in a lot of variation. We have no scientific based knowledge on allophonic rules, phonotactics or free variation of the language so far. In order to draw an – almost – clear border between synonyms and variants, we chose to treat signs with the same semantic content and variation in only one of the four parameters (hand shape, orientation, primary movement or place of articulation) as variants. Signs with two or more differences in form are treated as synonyms. The dictionary aims to be as descriptive as possible and therefore competing variants are shown side by side in the dictionary's entries.

5. Semantic features as lemmatization criteria

Lemmatizing is based only on phonology and semantics. We decided to leave out etymology, as no research has been done on the etymology of DTS. We allow only meanings that are semantically closely related from a synchronous point of view (as well as their transparent figurative uses) to occur together in one entry. Thus, strongly polysemous signs are often formally described as two or more homophone signs. For example, DTS expresses the meanings 'red' and 'social' through one manual expression, but the sign has two separate entries in the dictionary because the semantic

relation, although it might easily be explained diachronically, is considered synchronically opaque.

Some signs have the same phonological form due to closely related equivalents in Danish, *e.g.* the signs PRÆMIE 'reward' and PREMIERE 'opening night' with the Danish equivalents *præmie* and *premiere* which are almost homophones. Due to the semantic criteria each sign gets an entry in the dictionary. In all cases of homonymy, cross-references are given to the relevant homonym entries. Keeping a rather strict line of lemmatizing from semantic features contributes to the purpose of keeping the dictionary independent of the target language (Danish). This approach requires a thorough semantic analysis of every sign.

6. Concluding remarks

In the Danish Sign Language Dictionary we chose to represent sign lemmas by glosses, photos and video clips in order to make it as fast and easy as possible for all the potential users to read the sign headwords. For the purpose of searching and ordering the signs, their phonological features have been provided in the dictionary data, but this information is not shown explicitly to the user.

We lemmatize by the manual features of the sign's form and do not include mouth movements – in order to prevent an adoption of the target language's way of separating concepts in real life, and in order to prevent that polysemous signs with many possible mouth movements appear in the dictionary as large homonym groups.

We lemmatize by semantics in a way that only allows semantically closely related meanings in the same entry, in order to keep the dictionary as independent of the target language as possible when dealing with two languages which are – although quite different even in media – sociolinguistically very closely related, due to the fact that the speakers share a common political, social and cultural environment.

References

KLIMA, E. and BELLUGI, U. (1979). *The Signs of Sign Language*. Cambridge MA: Harvard University Press.

KRISTOFFERSEN, J.H. and NIEMELÄ, J.B. (2008). How to describe mouth patterns in the Danish Sign Language Dictionary. In R. Müller de Quadros (ed.). *Sign Languages: spinning and unraveling the past, present and future. TISLR9, forty five papers and three posters from the 9th Theoretical Issues in Sign Language Research Conference. Florianopolis, Brazil, December 2006*. Petropolis: Editora Arara Azul: 230-238.

LIDELL, S.K. and JOHNSON, R.E. (1987). The phonological base. *Sign Language Studies*, 64: 195-277.

PRILLWITZ, S., LEVEN, R., ZIENERT, H., HANKE, T. and HENNING, J. (1989). *HamNoSys. Version 2.0; Hamburg Notation System for Sign Languages. An introductory guide*. Hamburg: Signum.

STOKOE, W.C., CASTERLINE, C.C. and CRONEBERG, G.C. (1965). *A dictionary of American Sign Language on linguistic principles*. Washington: Gallaudet College Press.

Free online dictionaries: why and how?

Vincent Lannoy[1]
Ingénierie Diffusion Multimédia (IDM) – www.idm.fr

Abstract

New players in the field of online dictionaries, and more generally speaking, the specificities of the digital paradigm are challenging how publishers write, shape, label and market dictionaries. For long the question of the business model, distorted by the will to approach the web with products designed for print sales, has diverted publishers from the genuine advantages of lexicographical content to attract massive online audience. Moreover, the use of search engines in various languages by worldwide users is driving online dictionary publishing toward simultaneously a wider and a more accurate use of lexicography.

Keywords: free online dictionary, dictionary content, search engine optimization, 'pure players', local market, bilingual content, web traffic analysis tools.

1. Introduction

Free online dictionaries are very popular with users, who welcome the range of new possibilities they offer, among which multiple access routes, hyperlinking to other sites and incorporation of sounds (*cf.* Campoy 2002). For dictionary makers, however, the situation is very different. As pointed out by Dean (2005), "the dictionary market, thanks to the internet and the proliferation of home computing, is facing the biggest challenge yet to its previously untouched status". This "seismic shift in the availability of dictionary content" has begun to affect the sales of printed editions, a phenomenon which Kilgarriff (2010) describes as follows: "publishers who pay people (journalists, or lexicographers) to create content are losing out to others who recycle and re-use content, and make it available for free, undermining the publishers' income stream". As pointed out by Hutchins (2009: 15) in connection with online machine translation (MT) systems, "we know very little (indeed almost nothing) about who uses online MT and what for". This report aims at filling this gap in relation to online dictionary use in general. Making use of powerful web traffic analysis tools such as Google Analytics, we investigate users' queries with a view to optimizing publishers' online dictionaries.

[1] lannoy@idm.fr

2. Why free is a leading model for online dictionaries?

Dictionary books have built a genuine identity over the years: lexicographers work for renowned publishers according to specific rules and process; distribution channels are well organized and efficient, delivering for educational or public markets. The breakthrough of new actors, exclusively focused on the web, is a major upheaval as they deliver large corpora to a worldwide audience. Those "pure players" are now dominating the online dictionary market not only in terms of audience but also by establishing their brands, capable of competing with long-existing "brick and mortar" publishing houses.

We – as a technology provider – have entered this market to provide publishers with the means to compete with the most successful free websites: well indexed content delivered fast, search engine optimized pages, multi-lingual interfaces and extreme attention given to the data.

Those free online grounds offer a real opportunity for copyright owners, in control of their markets, that publishers are. Our experience, built by managing daily several major free online dictionary websites for English publishers, demonstrates the strong attraction power of dictionary content. A monolingual free dictionary website is approximately 400,000 pages to be indexed by search engines. Such a website, once properly optimized for search engines, is accessed monthly by 120,000 different keywords typed in search engines. In other words, by embracing a very large spectrum of the language and making it available for free on the Internet, a dictionary website is accessed by very diverse means, expressing different interests for the language and levels of expectations in terms of product delivered. How can a *closed* Premium website with a maximum of 50 free marketing pages offer a comparable visibility?

Quality of the content and publishers' care over data is playing a key role in building user loyalty and depth of visit on the website. On average, in a language learning context, we experience a visit to last between 5 and 7 pages, providing the publisher with the opportunity to be in contact with users for several pages. The question now is *to do what?* Most of the dictionary websites are dead-ends for the moment: a user enters for one or several definitions and leaves although his needs or interests can be much deeper: course books, vocabulary lists or exercises for learners, novels, reference content, targeted news for advanced users, etc. Affiliation models help propose not only the publisher's own content but complementary contents, products or services coming from partners. We are currently successfully experiencing with a partner the efficiency of the up-sales model based on free dictionary entries. Dictionary content is not only an efficient attraction point but plays also the role of a *user qualification filter for targeted up-sales*: dictionary is an intermediary between a query and a targeted product.

This selling function is to be duplicated and strengthened by targeting local markets. The most successful free dictionary websites in terms of audience are currently

proposing bilingual contents. This offer plays a key role on many markets, logically much wider than the English speaking sphere.

Now that we have quickly illustrated some specificities of the online dictionary market, let's detail the opportunities it offers in two areas:

Search Engine Optimization (SEO): why dictionary content is a marvellous material to answer a wide range of users' queries in search engines such as Google, Bing, Yandex or Baidu?

Local markets approach through bilingual content.

3. Why care so much about SEO?[2]

In most cases today, browsing the Internet consists in having a look at one's favourite websites (newspaper, e-mailing, social network) and searching for anything else into a search engine. To illustrate how prominent this usage is, consider the figure reported by a Hitwise survey:[3] Google Search was accounted to be 30% of the total UK web traffic in May 2008.

A search in Google.com to know what a dictionary is yields nearly two hundred thousand results among which, more than the first hundred results – it is difficult to assess accurately afterwards – provide free dictionary content.[4] Now, if one amends this query – [english dictionary], [online dictionary], [free dictionary], [dictionary], [spanish dictionary] – the result lists are changing, promoting websites in a different order.

The selection of pages and the order in which they are sorted out depend exclusively on each search engine algorithm. The capacity of each website to promote its own pages for a given search query (named *keyword* hereafter) over the competition is named Search Engine Optimization or SEO. Many criteria, specific to the dictionary world, are extraordinarily interesting.

3.1. Dictionary is porous to a large number of interests[5]

Let's take the example of a monolingual English dictionary of approximately 40,000 entries. Over a month, users enter the website by querying for 100,000 different keywords in search engines such as Google, Yahoo, Bing, varying from very generic – [dictionary] for instance or the product brand name – to accurate phrases and resulting in users landing on 35,000 different pages.

---

[2] The language used in this section for the case studies is mainly English as a convenient means to share examples.

[3] Read abstract here: http://www.hitwise.com/news/uk200805.html

[4] http://www.google.com/search?hl=en&q=dictionary

[5] The figures given in this section come from a traffic analysis tool (Google Analytics) monitoring the activity of a free dictionary website under our technical responsibility and from Google Webmaster Tools.

This figure can be compared with an educational Premium website launched by the same publisher. Over the same period of time, the website is accessed by approximately 500 different keywords in search engines, making visitors land on 35 different pages.

Crossing those figures illustrates a strong characteristic: content written for marketing purposes – like the free pages describing a Premium product – is very efficient to attract targeted traffic but cannot compete with the amazing acquisition capacities of a free dictionary, capable of addressing a wide range of the language used in search engine queries.

3.2. Dictionary content to prospect online[6]

Queries typed in search engines and leading to the free dictionary website can be split in two categories:

Branded queries, such as the website name or the publisher's brand name spelled in many ways;

Unbranded queries: any other queries like "dictionary" or "definition of something".

Figure 1 illustrates the number of visits coming from search engines to a monolingual English dictionary, broken down in those two categories over a period of time where SEO has been progressively enhanced.

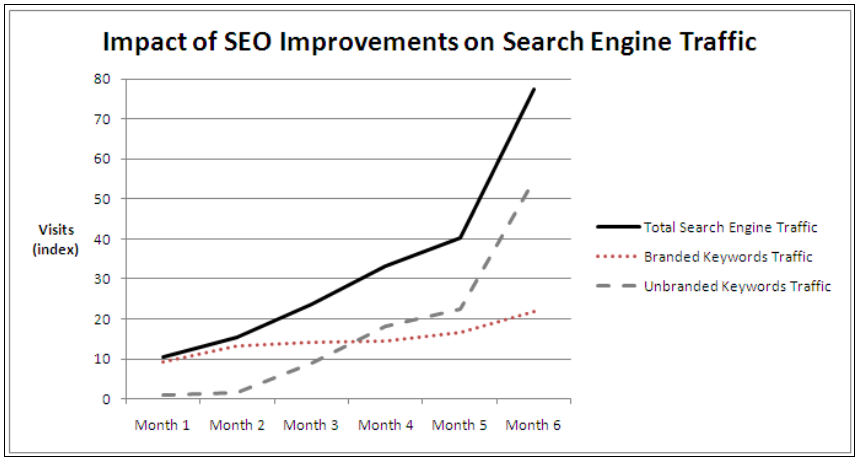


Figure 1: Impact of SEO Improvements on Traffic

Three main releases reflect different targets:

Month 1 = 1st release. Communication is focused on the website homepage, *i.e.* on the brand to be established online. There is no strong importance given to pages behind

[6] The figures given in this section come from a traffic analysis tool (Google Analytics) monitoring the activity of a free dictionary website under our technical responsibility, taken as a reference.

the home. Consequently, only the branded traffic reveals the website to end-users, meaning that those users already know the brand name. It is an already brand-sensitive audience coming to the website.

Month 2 = 2nd release. Second level of SEO to help search engines browsing the entry pages. More and more users enter the website by searching accurate language queries. The unbranded queries traffic is progressively larger than the branded one: the website is more and more visible on the web, via more and more queries landing on more and more pages. SEO is now helping prospecting new visitors.

Month 5 = 3rd release. A major SEO release targeted on optimizing definitions and phrases, *i.e.* the dictionary content rather than the brand name. The unbranded traffic is soaring, driving up the total website traffic and resulting also in an increase of the unbranded traffic. SEO shapes a virtuous circle where unbranded reputation ends with a better visibility of the brand.

Over the period, the share of new visitors coming from search engines has risen from 25% to more than 70%. SEO is definitely a power tool to prospect new users.

It is a quality of a free dictionary website to be able to generate such a large incoming traffic from unbranded queries. The capacity of the publisher to address the use of the language in its publishing and SEO policies shapes its online visibility and popularity.

3.3. Interest for phrases, an opportunity for publishers[7]

In the previous section, we have demonstrated the role of unbranded queries to drive traffic to a free website. But what are those queries, in particular when applied to a dictionary?

Some of the unbranded queries very logically contain the standard keywords *dictionary* or *English*. However, early players on the free dictionary market have a stronghold at the top of the search engines result lists for such standard queries.

Now, searches for English phrases like [*flushed with a howling success*], [*see to the children's breakfast*] or [*humanitarian grounds definition*] for instance, yield very different result lists, leaving room for websites with a smaller audience but with carefully edited and search engine optimized data. In the free dictionary website benchmark index we use in this presentation, 90% of the visits generated by unbranded queries are phrase centric instead of being worded around generic *dictionary* keywords. However, it is important to underline that we cannot identify demand for single word definitions (like [flower definition]), our benchmark index being too largely outscored by pure players.

Searching for phrases is a huge demand of the market and learners in particular. Moreover, phrases – by requiring specific linguistic and editing skills – help

[7] The figures given in this section come from a traffic analysis tool (Google Analytics) monitoring the activity of a free dictionary website under our technical responsibility.

differentiate publishers' content on the market. It is worth investigating ways of providing users with phrases and optimizing their indexing by search engines.

3.4. SEO for dictionaries is a specific know-how

Figure 2 compares total traffic of the website with number of visits brought by search engines. The SEO activity detailed in section 3.2 makes the total traffic grow. In other words, growth of the total website audience directly depends on the success of SEO.

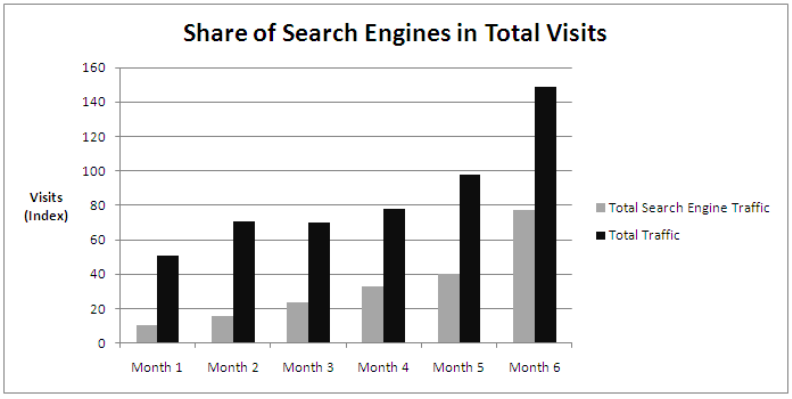


Figure 2: How SEO Feeds Traffic Growth

Search Engine Optimization work is to be thought and done in two distinct rounds:

On data first: what elements to value for search?

On the end-user application then, *i.e.* the website pages: how to position each page to provide clarity to end-users?

4. Local languages and local markets

To learn a language different from one's mother tongue or to understand a foreign language, people need good bilingual content to root their own translation capacities. This capacity to bridge the gap between two different languages is of major interest. And again Internet both strengthens and reveals this need. Let's ponder to what extent.

4.1. Native *vs*. English: diversity, not supremacy[8]

Internet users search more and more in their native language. This statement may sound either obvious or debatable, but it can be taken as a fact based on statistical data.

[8] In the current section and the followings, we use Google Insights for Search (http://google.com/insights/search/#). This tool makes it possible to compare volume of searches for given search queries over time, worldwide or on selected countries.

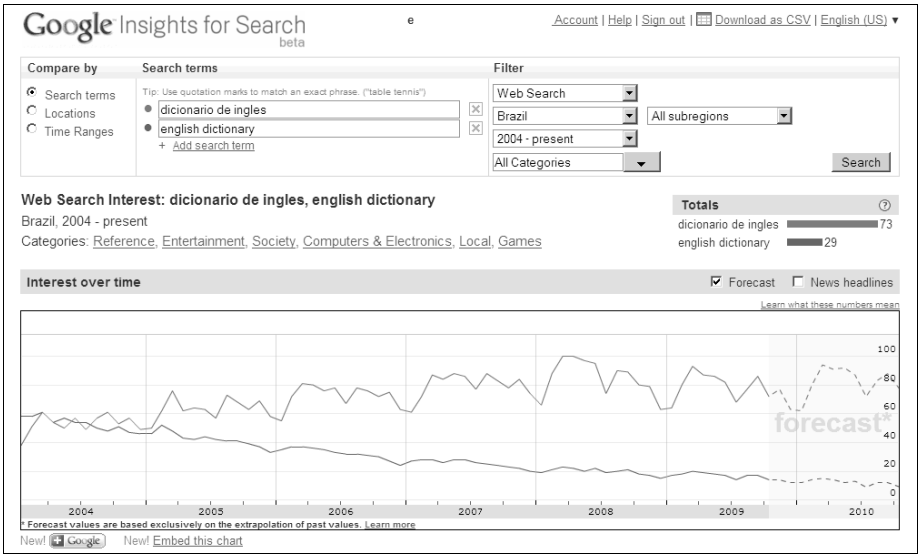


Figure 3: Comparison of volume of searches in Brazil for [english dictionary] and [dicionario de ingles]

Figure 3 compares the number of times ‘dicionario de ingles’ has been searched in Google from Brazil (upper line) with the number of searches for ‘english dictionary’, in Brazil too (lower line), from 2004 to 2009. One thing appears clearly: Brazilian Internet users are searching more and more for the same information by typing it in Portuguese rather than in English.

We take here the example of the Brazilian market but the same statement can be made, with only discrepancies in volumes but not in trends, on other markets such as Japan, Italy, France or Spain. It is the case in China too but more balanced, but not in Germany. Such a statement invites us to consider each market on its own, which is exactly what publishers have done for years with printed editions!

The stake for a publisher is to make their content visible in several targeted languages, even if the considered content is written in another language. Basically, to translate the gate towards its content.

4.2. To grow, provide bilinguals!

Another major result is that users are searching for bilingual content and this demand is increasing quickly.

In Figure 4, contrary to Figure 3, the line chart does not graph volumes of searches but growth rate of the search queries, relative to the growth rate of a given Google category of queries. In this case, we position our two queries within the Google category reference which contains all the reference works. The dark middle line represents the average growth rate of searches within the reference category. The

upper line stands for the ‘hindi english’ query when the lower line stands for the ‘english dictionary’ one. Those two queries must be read against the reference line.

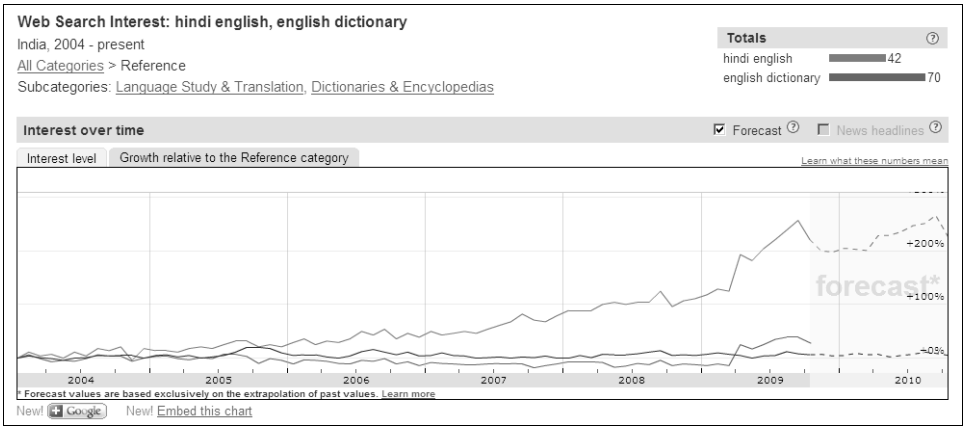


Figure 4: Comparison of searches growth rate in India between [hindi english] and [english dictionary]

The conclusion we can draw from this graph is clear: demand for Hindi English bilingual content is growing fast while perspectives of demand for monolingual English content are flat. As for the previous section, this statement does not only apply for India but for many other markets.

5. Conclusion: localize the approach

Crossing the two previous statements leads to the same conclusion: *users search for bilingual content in their native language*. And this statement can be taken for granted, at least on identified markets. Figure 5 uses the same growth rate comparison feature of Google Insights for Search.

According to the graph, in rough volumes, which are visible in the Totals panel (top right corner of the screenshot), demand for Spanish-English bilingual content, queried in English, is still much higher than queried in Spanish – index of 58 versus index of 3 on the whole period 2004-2009. However, if we consider growth rates instead of volumes, demand for bilingual content queried in Spanish is growing very fast.

The logical conclusion that can be drawn is that Internet strategy needs to be designed market by market and be prepared to localize the approach by a) providing bilingual content and b) translating the interfaces to maximize SEO efficiency. This is not entirely new: some actors are already entering those local markets. But it is good to confirm the validity of choices and so many combinations are offered that there is room for many players.

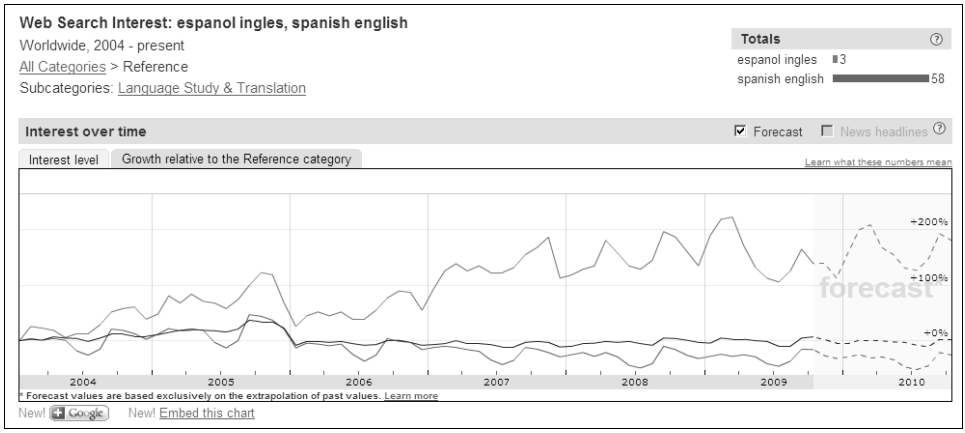


Figure 5: Comparison of searches growth rate worldwide between [espanol ingles] and [spanish english] queries

References

CAMPOY CUBILLO, M.C. (2002). General and specialised free online dictionaries. *Teaching English with Technology* 2/3. Retrieved from http://www.iatefl.org.pl/call/j_review9.htm on 14 February 2010.

DEAN, J. (2005). Lively legacy of 'dull work'. *Bookseller*. Retrieved from http://www.allbusiness.com/bookseller/20050415/4635798-1.html on 14 February 2010.

HUTCHINS, J. (2009). Multiple Uses of Machine Translation and Computerised Translation Tools. In *Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages – ISMTCL 2009*. Retrieved from http://www.hutchinsweb.me.uk/Besancon-2009.pdf on 14 February 2010.

KILGARRIFF, A. (2010). How to monetise a web presence (and hoover a moose). A report on the e-lexicography conference at Louvain-la-Neuve, 22-24 October 2009. Paper to be presented at EURALEX 2010.

The Hub&Spoke Model put into practice

Report on the semi-automatic extraction of a pre-version of a Finnish-Danish dictionary from a multilingual interlinkable database

Godelieve Laureys
Ghent University

Abstract

This paper deals with the results of a semi-automatic extraction of a Finnish-Danish dictionary from two existing bilingual contrastive databases, *viz.* Dutch-Danish and Dutch-Finnish. The process of linking and merging in line with the Hub&Spoke Model developed by Martin (2001) is described on the basis of an experiment and illustrated with examples. The possibilities of this semi-automatic extraction are discussed both from a practical and a theoretical point of view.

Keywords: bilingual lexicography, semi-automatic extraction, Hub&Spoke.

1. Introduction

In the period 1998-2009 two lexicographical projects leading to the Dutch-Danish (Laureys 2004) and the Dutch-Finnish/Finnish-Dutch (Laureys and Moisio forthcoming) bilingual dictionaries were successively conducted at the Department of Nordic Studies at Ghent University. Within the framework of these projects two electronic contrastive lexical databases have been elaborated, which the dictionaries are derived from. These databases contain, besides lexical data from the languages concerned, grammatical (*i.e.* morphological, categorial and relational) specifications and register-related (*i.e.* stylistic, geographical and pragmatic) information on the lexical entries and their examples. These specifications reach far beyond the scope of the graphic or even the electronic dictionaries and the annotated databases can therefore be exploited for further (contrastive) research.

Both the Dutch-Danish and the Dutch-Finnish dictionary rely on a corpus-based annotated Dutch lexical database, named Referentiebestand Nederlands (RBN). The RBN was commissioned by the Nederlandse Taalunie – The Dutch Language Union – and elaborated by an inter-university-consortium under the aegis of the Committee for Lexicographical Interlingual Resources (CLVV). The RBN is a shell database designed for the making of bilingual dictionaries with Dutch as a source language. It is a generic tool, which is not specifically oriented towards a particular target language.

Two concepts are vital for the understanding of the structure of the RBN: 'form unit' (FU) and 'lexical unit' (LU). Form units are strictly defined by categorial and morphological criteria and correspond to the orthographic representation of a given word. They make up the macrostructure. Lexical units refer to a specific semantic meaning denoted by a form unit. Each LU is marked by a semantic discriminator and the database also contains a short definition corresponding to the definition in a medium-sized monolingual dictionary. These definitions on the level of LU are not only an invaluable help for the editors, but they are also a guarantee for a univocal interpretation of the different readings of a Dutch entry. The microstructure of the RBN also comprises both canonical and contextual example units (EUs).

For the processing of the above-mentioned dictionaries we made use of the editor OMBI, which is tailor-made for RNB. OMBI is a quite advanced software tool, commissioned by the CLVV, which establishes translation-links between lexical units in L1 and L2. Bilingual dictionaries are to be processed by translating each LU into the target language. In the case of monosemy there is a virtual overlap between FU and LU, but in the case of polysemous entries this implies that the translation equivalents will be linked in the database on the level of the lexical units. OMBI is also equipped with a reversal function on the level of lexical units, which means that the reversing takes place at the level of meaning rather than form. With regard to polysemous form units this procedure implies that each meaning represented by a lexical unit is linked to a distinct lexical unit in the target language (Max 2007: 259-274).

For the Dutch-Danish dictionary project a lot of time and work was invested in establishing the Dutch macro- and microstructure and making an appropriate selection of the RBN examples. Actually, about 20% of the total project budget was spent on adjusting the RBN to our needs and purposes. In order to get a better return on investment it was decided to re-use this version of the RBN for future projects. For the Dutch-Finnish dictionary the RBN version elaborated for the Dutch-Danish dictionary, including the selections, modifications and additions made by the editorial team, was greatly re-used. This adapted version of RBN was re-imported into the OMBI editor and formed the Dutch basis for the new project. This means that the databases underlying the two dictionaries show a parallel structure as to the macro- and the micro-structure, *i.e*. distinction of meanings in polysemous words, grammatical and stylistic annotation and selection of examples.

2. The Hub&Spoke Model

The decision to re-use the RBN version was not only made for reasons of efficiency and cost savings. There was also an underlying meta-lexicographical vision as to the possibility of establishing a multilingual lexical database on the basis of several bilingual databases. The result of this working procedure is that we now have an excellent material, which allows us to put into practice the so-called Hub&Spoke model developed by Martin (2001, 2003, 2004). As illustrated in Figure 1 this model bears on the creation of an interlinkable multilingual database by merging two or more

compatible contrastive lexical databases which take their starting point in one and the same source language, the so-called "hub language". Within such a multilingual database new links can be established between entities of the L2-to-Ln-databases, the so-called "spoke languages", in order to establish a raw version of a bi-directional contrastive lexical database that could underlie a bilingual dictionary for a new language pair, *e.g.* Finnish and Danish. The linking of entities in L2-to-Ln (spokes) relies on the fact that the respective entities are linked to a L1 (hub)-entry in a univocal way. In our case, the link that should be established between a Finnish and a Danish LU in the merged database is traceable through the common link of both these LUs with a given Dutch LU. Therefore these links can be identified automatically.

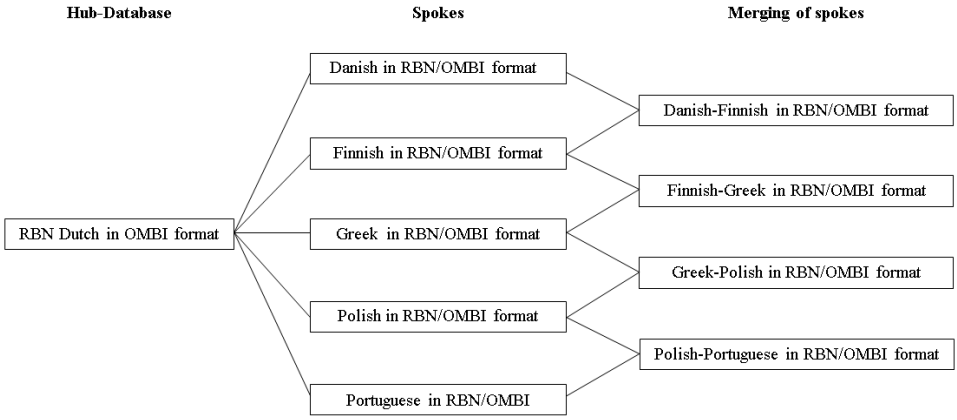


Figure 1: The Hub&Spoke Model

This model has been tested on an experimental basis by merging parts of the Dutch-Danish and the Dutch-Finnish database. By linking Danish and Finnish lexical units through their univocal and specified relation with a Dutch lexical unit, it was possible to extract an experimental draft of a Finnish-Danish dictionary. The semi-automatic linking of the two spoke-languages via the Dutch hub yields a workable pre-dictionary version, although it goes without saying that post-editing is necessary. In the following the experiment will be demonstrated with some examples, illustrating the benefits of the procedure as well as its limitations.

The Dutch-Finnish database is innovative in the sense that it has been elaborated as an integrated bi-directional database. The procedure implies that we have elaborated an integrated contrastive bi-directional database, from which both the Dutch-Finnish and Finnish-Dutch volumes can be derived. To this end all translation relations in the database have been marked according to their directionality status:

1. *bi-directional*, if the translation is univocal, adequate and relevant in both directions;
2. *unidirectional Du→Fi*, if the Finnish item only figures as a translation unit, but does not qualify as a LU or a EU in the Finnish-Dutch volume;
3. *unidirectional Fi→Du*, if the Dutch item shall be represented only as a translation unit in the Finnish-Dutch volume and be prohibited from appearing in the Dutch-Finnish direction.

This implies that all items in the database fall into three categories:

1. Dutch and Finnish items that should appear both in the Finnish-Dutch and in Dutch-Finnish volumes;
2. Finnish items that are coded to appear only in the Dutch-Finnish volume. These items are marked with <500>;
3. Dutch items that are coded to appear only in the Finnish-Dutch volume. These items are also marked with <500>.

3. Examples

The procedure can be demonstrated in the simplest way by selecting a monosemous Dutch word, which shows a one-to-one relationship with its translation equivalents. In this case the Danish and the Finnish translation can be linked to each other without any further complication, since they are linked to one and the same Dutch lexical unit. The example in Figure 2 shows how the Dutch LU *egel* 1. [Erinaceus europaeus] functions as a hub and the Danish and Finnish translation equivalents are the spokes.

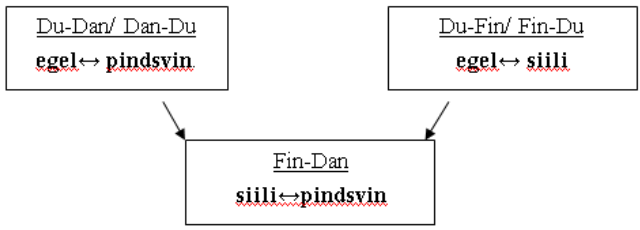


Figure 2: Monosemous one to one relationship.

In Figure 3, the monosemous Finnish word *opettaja* ‘teacher’, which is linked to the Dutch LU ‘leraar’ in the sense of “a person who teaches”, is linked to two different translation equivalents (FUs) in Danish. In the merged database it appears that *opettaja* is translated by both *lærer* and *underviser*, since both words are linked to the Dutch LU ‘leraar’: [10.] <NOUN> 1. [opetusta antava henkilö] *lærer, underviser*.

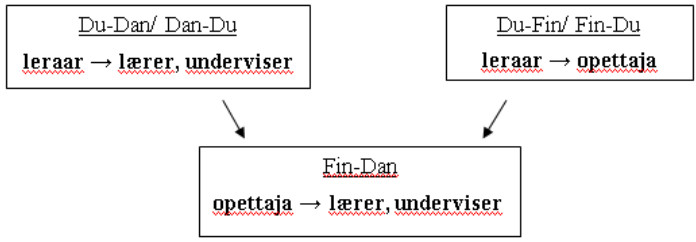


Figure 3: Monosemous one to two relationship.

A more complex issue is the linking of polysemous Finnish FUs with polysemous Danish FUs, especially when the matching involves more than one FU in the source or target language. In the example below we can see how the five meanings of the Danish word *bølge* correspond to two Finnish words *aalto* and *laine* and how the lexical units of *bølge* are redistributed as translation equivalents of *aalto* and *laine*. It should be kept in mind that the Finnish examples marked with the code <500> are just shown here to illustrate the procedure. They should be filtered out from a Finnish-Danish dictionary, since they were judged not to qualify for the microstructure of Finnish as a source language as mentioned above. Therefore they are crossed out below.

aalto [101.] [I] <NOUN>

1. [veden pinnan kohouma] *bølge* **a.** ~~*(500)* aallot~~ **b.** ~~*(500)* korkea aalto~~ **c.** korkeat aallot *høje bølger* **d.** ~~*(500)* raivoavat aallot~~ **e.** keinua aalloilla *ride på bølgerne* **f.** vaipua aaltoihin *forsvinde i bølgerne*

2. [värähdysliike] (fys) *bølge* **a.** pitkät aallot *langbølge* **b.** lyhyet aallot *kortbølge* **c.** lähettää lyhyillä aalloilla *NIHIL*

3. [kuv: aallon kaltainen ilmiö] *bølge* **a.** uusi aalto *NIHIL* **b.** vihreä aalto *grøn bølge* **c.** kuumat aallot *hedetur*

aalto: *English translation*

1. [elevation on the surface of water]: wave **a.** <500> **b.** <500> **c.** high waves **d.** <500> **e.** ride on the waves **f.** sink beneath the waves

2. [vibratory motion]: wave **a.** long wave **b.** short wave **c.** broadcast on short waves

3. [fig..]: wave **a.** the New Wave **b.** phased traffic lights/signals **c.** hot flashes

laine [48.] <NOUN>

1. [aalto] *bølge* **a.** valaa öljyä laineille *gyde olie på vandene*

2. [aalto muistuttava kuvio *bølge* **a.** ~~(*500*) jonkun hiukset ovat laineilla~~ **b.** (*500*) ~~kammata laineita hiuksiin~~ **c.** muotoilla hiukset laineille *lave bølger i håret*

laine: *English translation*

1. [elevation on the surface of water]: wave **a.** pour oil on troubled waters

2. [pattern resembling a wave]: wave **a.** <500> **b.** <500> **c.** have one's hair set/water-waved

The most complex instance is the case where a polysemous Finnish word is linked to LUs belonging to several polysemous Danish words. In that case the discrepancy between the lexical structures of the two languages is at its utmost – apart from the cases where no translation equivalent can be given, since a given meaning is not lexicalised in the L2. In the example below the Finnish noun *kello* can be split up in three different lexical units, respectively denoting 'a time measuring device' *(clock)*, a point of time *(time)*, a ringing device *(bell)*. These lexical units of one Finnish word (FU) correspond to three different words (FUs) in Danish: *ur, time, klokke*.

kello [101.] <NOUN>

1. [aikaa mittaava laite] *ur* **a.** katsoa kelloa *se på uret* **b.** kello lyö yksitoista *klokken slår elleve* **c.** kello edistää/on edellä *uret er foran* **d.** kello jätättää/on jäljessä *uret er bagud* **e.** kelloni jätättää/edistää kaksi minuuttia *mit ur er to minutter bagefter/foran* **f.** kello seisoo *uret er gået i stå* ~~**g.** (*500*) kello on pysähtynyt~~ **h.** vetää kello *trække uret op* **i.** siirtää kelloa eteenpään *NIHIL* **j.** asettaa kello oikeaan aikaan *stille uret* ~~**k.** (*500*) asettaa kellot samaan aikaan~~ ~~**l.** (*500*) kellon koneisto~~

2. [ajankohta] *time* **a.** kysyä kelloa *NIHIL* **b.** paljonko/mitä kello on? *Hvad er klokken?* **c.** ~~(*500*) voitteko sanoa, paljonko kello on?~~ **d.** ~~(*500*) kello on yksi~~ **e.** kello on viisi *klokken er fem* **f.** kello viideltä *klokken fem* **g.** tasan kello kaksitoista *NIHIL* **h.** nukkua kellon ympäri *sove i tolv timer*

3. [soittolaite] *klokke* **a.** soittaa (kirkon)kelloja *ringe med (kirke)klokkerne* **b.** kellot soivat *NIHIL* **c.** soittaa (ovi)kelloa *ringe på* **d.** ~~(*500*) soittaa kelloa *stemme dørklokker som leg*~~ **e.** kello soi *det ringer på døren* **f.** ~~(*500*) hänelle tuli toinen ääni kelloon~~ **g.** ~~(*500*) sitten hänelle tuleekin toinen ääni kelloon~~ **h.** nyt on toinen ääni kellossa *nu får piben en anden lyd* **i.** ei vahinko tule kello kaulassa *en ulykke kan ske hurtigt, det kan pludseligt gå galt*

kello: English translation

1. [time measuring device]: *clock.**a.** look at (consult) one's watch **b.** the clock strikes eleven **c.** the clock is fast **d.** the clock is slow **e.** my watch is two minutes slow/fast **f.** the clock has stopped **g.** <500> **h.** wind [up] the clock **i.** put the clock forward **j.** set the clock **k.** <500> **l.** <500>*

2. [point of time]: *time **a.** ask the time **b**. what time is it? **c.** <500> **d.** <500> **e.** <500> **f.** it is five o'clock **g.** at five o'clock **h.** at twelve o'clock sharp **i.** <500> **j.** sleep (a)round the clock*

3. [ringing device]: *bell* ***a.*** *toll the [church] bells* ***b.*** *the bells are ringing/tolling* ***c.*** *ring the doorbell,* ***d.*** *<500>* ***e.*** *the bell is ringing* ***f.*** *<500>* ***g.*** *<500>* ***h.*** *change one's tune, sing a different tune* ***i.*** *accident's will happen, accidents easily happen*

Another example is the Finnish word *virta*, which has no less than five different lexical units, some of which correspond to the polysemous Danish word *strøm*, while others have different translation equivalents in Danish.

virta [9.] [K] <NOUN>

1. [leveä joki] *flod*

2. [virtaava neste] *strøm* **a.** kyynelten virta *NIHIL* **b.** valua virtanaan *strømme i store mængder*

3. [virtaus] *strøm* **a.** siellä on kova virta *der er stærk strøm* **b.** virta vei veneen mukanaan *NIHIL* **c.** uida virtaa vastaan *svømme imod strømmen* **d.** mennä/ajelehtia virran mukana *følge med strømmen*

~~4.~~ [sähköstä] *strøm* **a.** ~~*(500)* tuossa johdossa kulkee virta~~ **b.** kytkeä virta *tænde for, starte, sætte igang* **c.** katkaista virta *slukke for* **d.** virta on poikki *strømmen er gået, der er strømafbrydelse* **e.** kuluttaa paljon virtaa *bruge meget strøm, sluge strøm <informal>* **f.** (*500*) ~~syödä paljon virtaa (informal)~~

5. [koko ajan etenevä joukko] *strøm, tilstrømning* **a.** autoja kulki jatkuvana virtana *der var en lind strøm af biler* **b.** (*500*) ~~uteliaiden virta~~

virta: English translation

1. [broad riveri]: river, stream

2. [flowing liquid]: stream, flow **a.** flood of tears **b.** pour down, flow in streams

3. [current]: current **a.** there is a strong current **b.** the boat was swept away by the current **c.** swim against the current **d.** swim with the stream

4. [electricity]: current **a.** <500> **b.** switch on the current **c.** switch off the current **d.** the power is off **e.** consume a lot of energy

5. [flow constantly moving forwards]: stream, flow, flood **a.** cars were running as a constant stream

From these instances of a Finnish-Danish pre-version we can easily see that quite a lot of the translations are generated automatically by the system, both at the level of the lexical units and that of the example units. In the cases where no translation was available in the system NIHIL is filled out. For these examples the underlying Dutch translation does not function as a hub, since it does not appear in the Dutch-Danish database. This can either be due to the fact that these examples have been altered or that they have been inserted by the Finnish editors when establishing the Finnish microstructure after the reversion of Dutch-Finnish. In fact not all translated example units, however correct they might be as translations of Dutch examples, suit as entries in the microstructure of Finnish headwords, seen from the source language perspective.

4. Evaluation

When we look at the outcome of the semi-automatic extraction we may conclude that no or only marginal post-editing is needed in a number of cases. First of all, as could be expected, this applies to the case of monosemous words. One should remember that this amounts in fact to c. 60% of the macrostructure. Secondly, the results are also optimal in the case of multi-word expressions and idioms, provided that there is a matching degree of lexicalization in both languages. Thirdly, the translation units of canonical examples can in many cases be linked without any further manual editing.

Some additional post-editing is needed in the case of polysemous nouns and verbs and especially in the case of adjectives, since the latter relate to the semantic and pragmatic features of the nouns they can define. Moreover, most of the contextual examples need some post-editing because of their semantic specificity and syntactic complexity.

Rather poor results are obtained in the case of high frequency nouns and especially in the case of verbs with a high degree of polysemy and one might ask the question whether a semi-automatic extraction is worth while. This is of course also true in the case of function words.

There remain a number of problems of a more systematic character that need a further investigation. Actually, it cannot be excluded that text strings in the spokes are to a certain extent biased by the hub. Contextual examples are a case in point and the bias might either concern the very selection of the example or the exact wording. During the post-editing vigilance should be exercised at this point. In spite of the frequent use of corpora and of the internet during the translation process it should also be kept in mind that the examples in the spokes are not genuinely corpus-based, but are the result of editorial translation.

Finally, a monitoring device is needed in order to screen the material as to the contrastive relevance of both the entries and the examples with respect to each new language pair. At the same time omissions must be detected and repaired.

5. Conclusion

On the basis of this experiment we may conclude that the derivation of a Finnish-Danish dictionary out of the Dutch-Danish and the Dutch-Finnish database is a feasible enterprise. This merged Finnish-Danish database could subsequently be reversed in order to derive a Danish-Finnish pre-version. Since both the Danish and the Finnish databases are conceived as bi-directional resources, all relevant grammatical (*i.e.* morphological) information as well as all stylistic and pragmatic labels are already in place for both spoke-languages. The semi-automatic linking of the two spoke-languages via the Dutch hub thus yields more than a raw pre-dictionary version. It goes without saying that post-editing is necessary, but still the amount of labour can be reduced drastically and the re-use of resources yields a good return on investment.

These experiments are interesting on the one hand in terms of practical lexicography, since they provide us with information about how to optimise lexicographical processes and on the other hand in terms of meta-lexicography, since they show which parts of the lexical database are best suited for automatic linking and yield the most operational results. Finally, these simulations may also shed new light on lexicalisation processes in different languages with regard to words, grammatical and pragmatic collocations and idiomatic expressions.

References

LAUREYS, G. (2004). *Hollandsk-Dansk ordbog*. Copenhagen: Gyldendals forlag (also published as: *Prisma Groot Woordenboek Nederlands-Deens*. Utrecht: Het Spectrum).

LAUREYS, G. and MOISIO, M. (forthcoming). *Prisma Groot Woordenboek Nederlands-Fins en Fins-Nederlands*. Utrecht: Het Spectrum.

MARTIN, W. **(**2001). Linkability, the Hub-and-Spoke Model and the (semi-)automatic derivation of a German-French Dictionary. In H.E. Wiegand (ed.). *Studien zur zweisprachigen Lexikographie VI, Germanistische Linguistik* 16. Hildesheim: Olms: 67-91.

MARTIN, W. (2003). Lexicography, Lexicology, Linking and the Hub-and-Spoke Model. In W. Botha (ed.). *Festschrift D. Van Schalkwijk.* Stellenbosch: Buro van die WAT: 237-249.

MARTIN, W. **(**2004). SIMuLLDA, the Hub-and-Spoke Model and Frames or How to Make the Best of Three Worlds? *International Journal of lexicography*, 17(2): 175-187.

MAX, I. (2007). OMBI: The Practice of Reversing Dictionaries. *International Journal of Lexicography*, 20: 259-274.

Entry menus in bilingual electronic dictionaries

Robert Lew[1], Patryk Tokarek
Adam Mickiewicz University, Poznań

Abstract

The study undertakes to assess the efficiency of entry menus in bilingual dictionaries in the electronic format. An experimental dictionary interface is tested for performance in terms of access speed and task success. The task underlying dictionary use is guided Polish-to-English translation, performed under three conditions by 90 Polish learners of English. The first version of the dictionary displays a complete polysemous entry immediately after an entry is selected. In the second version the user is presented with a menu of senses; once the user clicks on the sense of choice, the full entry is shown, scrolled to the selected sense. The third version is identical to the second, but, in addition, the target sense is highlighted. Our results indicate that a combination of menu-guided sense access and target sense highlighting is effective in terms of both speed and task success, at both user levels investigated. In contrast, the menu alone is not significantly more effective than presenting the full entry at once.

Keywords: bilingual dictionaries, dictionary access, entry navigation, sense-facilitating devices, entry menu, sublemmatic addressing.

1. Background

1.1. Access in electronic dictionaries

Efficient access to lexicographic data is one area where electronic dictionaries are expected to excel compared to traditional paper dictionaries (de Schryver 2003). For access to be efficient, users have to be able to find just the information they need (as long as the relevant data are in the dictionary), and they have to be able to complete the search quickly enough for it to be worth their while.

1.2. Problems with accessing dictionary senses

All too often, users fail to locate information in dictionaries even when the relevant lexicographic data is actually there (Nesi and Haill 2002). One particularly problematic step in dictionary consultation is the selection of the relevant sense in polysemous entries. Studies indicate that language learners will often stop at the first

[1] Department of Lexicology and Lexicography, School of English, Adam Mickiewicz University, rlew@amu.edu.pl

sense unless there is a clear indicator that this sense is not appropriate (Tono 1984; Lew 2004).

2. Sublemmatic access facilitators

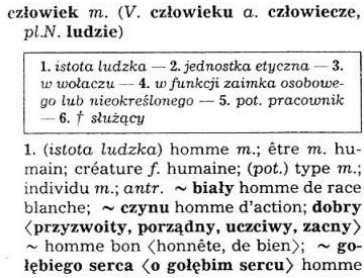
To remedy the above problem and facilitate quick and accurate access to the relevant sense, highly polysemous entries can be enriched with entry-internal (or sublemmatic) *access facilitators* which would guide the user to the likely sense at a glance. Two types of such devices have become rather well known from the recent editions of English monolingual learners' dictionaries, which have, one by one, started providing them in longer, highly polysemous entries.

2.1. Signposts

The first of these is a system of sense indicators given at the beginning of each sense. Depending on the dictionary publisher, they are variously referred to as *signposts*, *guidewords*, *shortcuts*, or *mini-definitions* (Tono 1997; Bogaards 1998; Lew and Pajkowska 2007).

2.2. Menus

The signpost system is based on brief sense indicators distributed across the specific senses; a distinct alternative is the entry menu. Here the idea is to gather sense indicators in a single block at the top of the entry proper, creating a kind of "table of contents" of the dictionary entry. An early example from a large Polish-French dictionary published in Poland in 1983 (*Grand dictionnaire polonais-français. Wielki słownik polsko-francuski*) is shown in Figure 1.



człowiek *m.* (*V.* **człowieku** *a.* **człowiecze**, *pl.N.* **ludzie**)

1. *istota ludzka* — 2. *jednostka etyczna* — 3. *w wołaczu* — 4. *w funkcji zaimka osobowego lub nieokreślonego* — 5. *pot. pracownik* — 6. † *służący*

1. (*istota ludzka*) homme *m.*; être *m.* humain; créature *f.* humaine; (*pot.*) type *m.*; individu *m.*; *antr.* **~ biały** homme de race blanche; **~ czynu** homme d'action; **dobry** ⟨**przyzwoity, porządny, uczciwy, zacny**⟩ ~ homme bon ⟨honnête, de bien⟩; **~ gołębiego serca** ⟨**o gołębim sercu**⟩ homme

Figure 1. Example entry menu from a Polish-French dictionary

As seen in the example, entry menus in this dictionary are entirely in Polish and appear to be specifically addressed to the Polish user engaged in foreign language text production or L1→L2 translation. The menu presents a list of sense numbers followed by semantic or grammatical function indicators in the source language, and also some

domain, register and currency labels. Note, however, that the menu does not provide any target language equivalents, and this appears to be a good decision if we consider the dangers of the user stopping the consultation at the menu itself, and never going to the full entry. In contrast, menus in a companion French-Polish volume mostly consist of translation equivalents. In this section of the dictionary the target language is Polish, so the overarching principle seems to be to present menus in the native language of the typical user. This makes sense: a menu is for scanning, and it is obviously easier to scan text in your native language. The dangers of users grabbing Polish equivalents from the menu itself and never reading the full treatment are less severe in decoding than they are in encoding.

2.3. Are entry menus effective?

The use of entry menus to facilitate entry navigation was suggested by Yukio Tono (1984). In a follow-up study (Tono 1992; updated version in Tono 2001, Chapter 10), Tono tested the idea on Japanese learners of English and found the menu helpful in assisting the process of sense selection for learners at the level of junior high school, but observed no such effect for the more advanced group of college students. Tono concluded that the difference was due to poorer reference skills in the junior group. English-Japanese entries were used with invented headwords; the outcome measure was the accuracy of sense selection. Access speed and translation accuracy were not measured.

Apart from Tono's study, there have been a small number of studies looking at the role of signposts for access speed and accuracy (Tono 1997; Bogaards 1998; Lew and Pajkowska 2007), but their results are not entirely consistent and their relevance for menu-equipped entries is rather indirect, so we shall not summarize them here.

2.4. Entry menus in electronic dictionaries

The idea of entry menus in electronic dictionaries can be traced back to their paper ancestry. For example, the *Macmillan English Dictionary*, known for its principled application of entry menus since its first edition (Rundell 2002), has carried over the system to both the PC and online[2] versions. The same is true of electronic versions of the *Longman Dictionary of Contemporary English* (Mayor 2009).

But, quite apart from the lexicographic tradition, the concept of a menu as such is a very familiar one in IT: the average computer users can be expected to be fairly accustomed to using menu-driven interfaces to find their way through collections of options that would be hard to take in if presented all at once. Similar rationale compelled Hulstijn and Atkins (1998: 16) to contrast the following access routes (I leave out the third alternative here):

[2] http://www.macmillandictionary.com/

1. The whole entry is simultaneously available (as it is in a normal paper dictionary).
2. The information in the entry is presented in various phases. At each step, users are given two or more options to choose from, and are thus led towards the information they will finally select (whether correct or incorrect), without seeing all the rest of the information which the entry contains.

Hulstijn and Atkins do not use the word *menu*, and if we were to follow option 2, then senses other than the one thought to be relevant would actually be suppressed. However, this is not what typically happens in today's e-dictionaries and it is interesting to consider the reasons. These might be any combination of the following: a desire to maintain a degree of continuity in lexicography; a lack of confidence in the user being able to competently select what they really need; a realization that dictionary senses are not in fact discrete entities and the entry is a cohesive text of sorts; not least, inertia and technical difficulty might play a role.

The entry menu may be always-on (*MED*) or invoked by request only (*LDOCE*). Typically, senses on the menu are clickable and the display will ideally scroll to position the target sense at the top of the window. However, depending on the particular combination of sense position, page layout, window size and magnification, the display may not scroll reliably. For example, if the target sense is found towards the bottom of the browser page and the dictionary is viewed in a maximized window on a large display, our sense may not make it to the top of the window.

3. The study

3.1. Purpose

Given the scarcity of experimental studies of entry menus in general, and a complete lack of such studies in electronic dictionaries, we wanted to test the usefulness of entry menus as sublemmatic access facilitators in online bilingual dictionaries. In addition, we elected to compare a new presentation device afforded by modern advances in information technology: the highlighting of the target sense, in the hope that such highlighting would help users locate the target sense more quickly and more accurately. Finally, we wished to see whether subject level or entry length made any difference.

3.2. Instruments and subjects

In order to meet our objectives, we created an experimental electronic Polish-English dictionary interface in HTML and PHP[3], in three versions. In this implementation a list of headwords was presented as an alphabetical list on the left, as is common in

[3] We wish to thank Mr Michał Katulski for his help with designing the interface.

electronic dictionaries. But when one of the Polish headwords was selected with the mouse, the three versions of the dictionary behaved differently:

1. the complete entry is presented at once (the NO MENU condition);
2. a clickable entry menu was displayed, and upon clicking on a specific sense, the complete entry was displayed, scrolled to the selected sense (the MENU condition);
3. an entry was displayed exactly as in 2. above; when the user selected a sense, the complete entry was displayed, again scrolled to the selected sense, but in addition the selected sense was highlighted (the MENU+HIGHLIGHTING condition).

The experimental dictionary featured twenty entries relevant to the task, ten nouns, nine verbs and one adjective, all fairly common words, but here used in the less common and less obvious senses. The entries were adapted from a leading Polish-English dictionary (Linde-Usiekniewicz 2002). They varied in length between four and twelve senses each.

Our subjects were 90 Polish learners of English aged between 16 and 19, at two levels of secondary school (*Liceum*) corresponding to two proficiency levels which we will here refer to as *Lower* and *Higher* for convenience:

1. 1st grade students (pre-intermediate, or A2 according to the Common European Framework of Reference (CEFR));
2. 3rd grade students (intermediate, or B1 according to the CEFR).

All subjects were computer literate.

3.3. Procedure

Before the experiment proper, the complete procedure was piloted on five intermediate-level students other than experimental subjects. The pilot study revealed a few minor (mainly technical) problems which were corrected for the main experiment.

Subjects completed the experimental task individually at the office of one of the researchers. They all used the same computer and the Opera 9 browser. They were assigned a guided Polish-to-English translation task to be completed on paper. The task consisted of twenty partially translated sentences with gaps at problematic lexical items. Students were instructed in Polish to use the online dictionary to look up those items and complete the translations. The dictionary version was assigned randomly.

3.4. Data analysis

All of the subjects' consultation activity was logged in files, including time stamps accurate to within 1ms. Their translations were later examined and translation accuracy scores were calculated on that basis.

A 3-way ANOVA was computed on sense access times and translation accuracy scores, with proficiency level and dictionary version as between-subjects factors, and entry length as a within-subject factor. Where called for, post-hoc analysis was conducted using the Tukey Honest Significant Difference formula.

4. Results

Figure 2 presents the mean per-entry access times for each of the three dictionary versions. It will be seen that subjects using the version with target sense highlighting took the least time, 25.6 seconds on average, as against 34.1 seconds in the no menu version, and 33.2 seconds for those using the menu but with no highlighting. This difference is highly significant ($F_{(2,1898)}=58.4$, $p<0.001$). Post-hoc analysis reveals that access times are significantly shorter in the menu with highlighting dictionary version than for each of the other two versions.

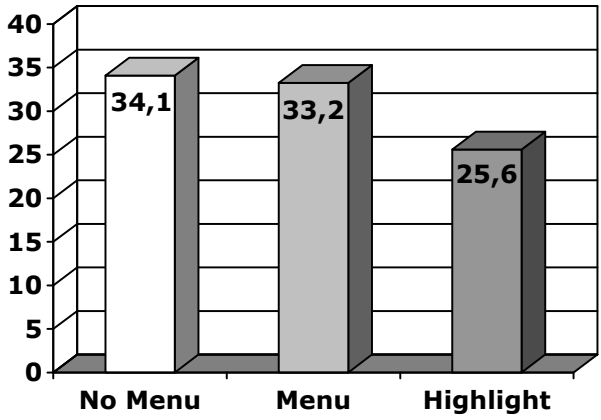


Figure 2. Mean sense access time (in seconds per entry) for the three dictionary versions

Next, let us examine how the three dictionary versions serve the two proficiency levels. Figure 3 presents the interaction of version and level on mean access time.

For the lower-level students there is a neat stepwise progression from the no-menu version, which takes the longest (37.7 seconds on average), through the menu version (33.3 sec), to the menu with highlighting, which is the fastest (28.0 sec). In the higher-level group the menu-with-highlighting version is again the fastest (23.1 sec), but here the bare menu performs worse than the no-menu version. Comparing the two proficiency levels using the same interfaces, higher-level students get to their senses faster than lower level students when working with the no-menu version as well as the menu-with-highlighting version, but with the bare menu there is no difference between the higher- and lower-level users. It looks then as if the higher-level students — but

not the lower-level ones — somehow get confused by the bare menu version. One possible explanation for this somewhat paradoxical effect might be that our third-graders are already in possession of established habitual reference routines, and these are thwarted when facing a not-too-familiar element: the entry menu. In contrast, their younger counterparts may be less set in their ways when it comes to dictionary use. We have to stress, though, that even if this is a reason for the higher-level users performing worse with menus, such an undesirable effect is fully compensated by the addition of target sense highlighting. It seems, then, that you cannot lose with highlighting.

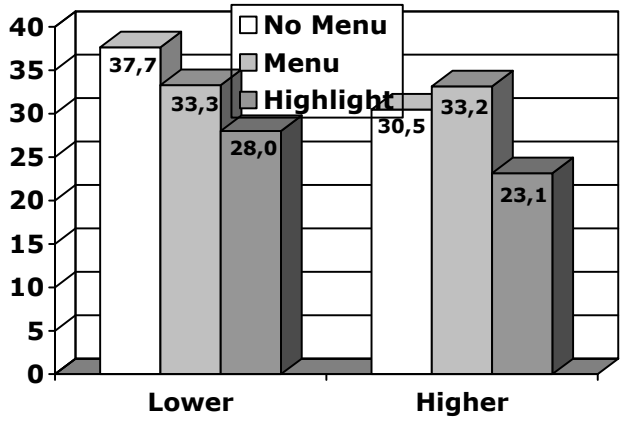


Figure 3. Mean sense access time by version and proficiency level

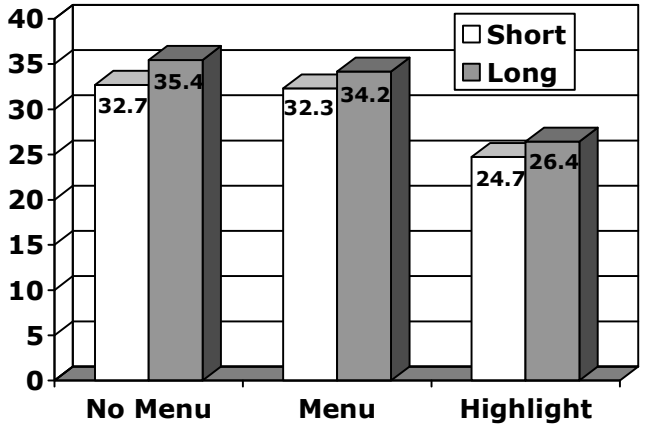


Figure 4. Mean sense access time by dictionary version and entry length

Next, let us examine the effect of entry length on lookup time. Recall that our dictionary entries consisted of between four and twelve senses. We classified entries of between four and six senses as short, those of seven and more senses as long. With this distinction in mind, refer to Figure 4, which plots mean sense access time for the three dictionary versions, broken down by entry length.

For all three dictionary versions, longer entries take longer to consult, as would be reasonable to expect. But it is striking just how stable this difference is across the three dictionary versions. Clearly, it makes no difference which version you use: a longer entry just takes a fraction longer to process. This makes good sense: even if you include entry menus, the menus themselves must be longer for the longer entries. The interaction of dictionary version and entry length is not significant (two-way ANOVA, $F_{(1,10)}=0.041$, p=0.8, n.s.).

Finally, let us examine the task-related variable, that is translation error rates for users working with the three dictionary versions, again broken down into the two proficiency levels. The respective rates are presented in Figure 5.

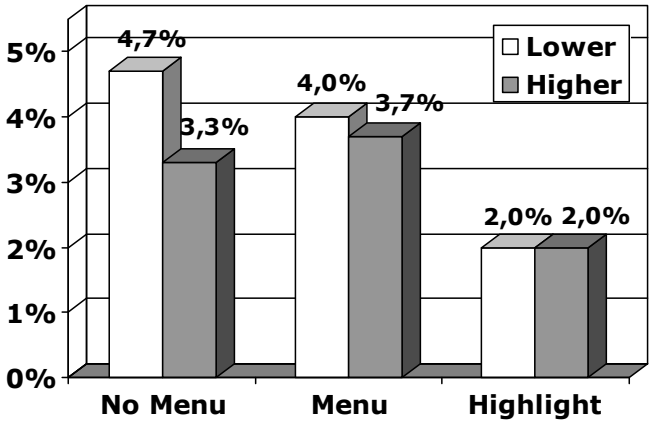


Figure 5. Task error rate by version and level

The error rate for the highlighting-equipped entries is halved compared to the other two interfaces. The effect of dictionary version does not reach significance at the 5% level (one-way ANOVA, $F_{(2,87)}=2.6$, p=0.08), but the power of the test appears to be somewhat compromised due to a floor effect. However, we should note that reducing the error rate by half would clearly represent a marked improvement, and thus not a trivial finding, even if statistically marginal. Looking at the two less successful versions, we may note that whereas the higher-level students predictably perform better on the translation task than the first-graders when working with the no-menu interface, their advantage just about evaporates when using the bare menu. Here again, we see a problem with the third-grade students interacting with the menu.

5. Discussion and conclusion

Our results indicate that the advantage in access speed comes from sense highlighting rather than the presence of an entry menu alone. However, when the level of proficiency is factored in, bare menus appear to facilitate access for lower-level students, but hinder higher-level users. In contrast, menus with highlighting seem to assist users at both levels in equal measure.

Translation error rates are largely unaffected by the presence of menus alone, but are reduced by half when highlighting is added.

Thus, we can conclude that in online bilingual dictionaries (used in guided production tasks) target sense highlighting is an effective technique, offering significant benefits beyond the bare menu, both in speed and accuracy, and seems to work well for both levels examined. In contrast, the menu alone is not very helpful, and is actually counterproductive to higher level students in our sample.

The recommendation that follows from our study is that target sense highlighting is a navigation device worth including for polysemous entries, as it assists users in reaching the relevant sense more quickly, and with fewer errors. Before electronic dictionaries get intelligent enough to "guess" which sense is actually needed, they can help users navigate the entry better with sense highlighting, thus contributing to a more user-friendly interface.

References

BOGAARDS, P. (1998). Scanning long entries in learner's dictionaries. In Th. Fontenelle, Ph. Hiligsmann, A. Michiels, A. Moulin and S. Theissen (eds). *Euralex '98 actes/proceedings*. Liège: Université Départements d'Anglais et de Néerlandais: 555-563.

DE SCHRYVER, G.-M. (2003). Lexicographers' dreams in the electronic-dictionary age. *International Journal of Lexicography*, 16(2): 143-199.

Grand dictionnaire polonais-français. Wielki słownik polsko-francuski. (1983). Warszawa: Wiedza Powszechna.

HULSTIJN, J.H. and ATKINS, B.T.S. (1998). Empirical research on dictionary use in foreign-language learning: Survey and discussion. In B.T.S. Atkins (ed.). *Using dictionaries. Studies of dictionary use by language learners and translators*. Tübingen: Niemeyer: 7-19.

LEW, R. (2004). *Which dictionary for whom? Receptive use of bilingual, monolingual and semi-bilingual dictionaries by Polish learners of English*. Poznań: Motivex.

LEW, R. and PAJKOWSKA, J. (2007). The effect of signposts on access speed and lookup task success in long and short entries. *Horizontes de Lingüística Aplicada*, 6.2: 235-252.

LINDE-USIEKNIEWICZ, J. (2002). *Wielki słownik angielsko-polski PWN-Oxford*. Warsaw: PWN.

MAYOR, M. (ed.) (2009). *Longman dictionary of contemporary English*, 5th edition. Harlow: Longman. (LDOCE5).

NESI, H. and HAILL, R. (2002). A study of dictionary use by international students at a british university. *International Journal of Lexicography*, 15(4): 277-305.

RUNDELL, M. (ed.) (2002). *Macmillan English dictionary for advanced learners*. Oxford: Macmillan Education. (MED1).

TONO, Y. (1984). *On the dictionary user's reference skills*. B.Ed. Thesis, Tokyo Gakugei University, Tokyo.

TONO, Y. (1992). The effect of menus on EFL learners' look-up processes. *Lexikos*, 2: 230-253.

TONO, Y. (1997). Guide word or signpost? An experimental study on the effect of meaning access indexes in EFL learners' dictionaries. *English Studies*, 28: 55-77.

TONO, Y. (2001). *Research on dictionary use in the context of foreign language learning: Focus on reading comprehension*. Tübingen: Niemeyer (*Lexicographica Series Maior* 106).

Designing specialized dictionaries with natural language processing

Examples of applications in the fields of computing and climate change

Marie-Claude L'Homme[1]
OLST, Université de Montréal

Abstract

This article examines some of the changes brought about by the introduction of natural language processing (NLP) technologies in terminology work. During the last decades, terminology projects have changed drastically due mostly to the introduction of computer applications and the availability of corpora in electronic form. The basic tasks related to specialized dictionary compilation are listed and examined from the point of view of their possible automation. Examples of applications to the special subject fields of computing and climate change in English and in French serve to illustrate the process of dictionary compilation with NLP technologies.

Keywords: specialized dictionaries, terminology, natural language processing.

1. Introduction

During the last decades, terminology work has changed drastically due mostly to the introduction of computer applications and the availability of corpora in electronic form. Although the main steps of the methodology have remained basically the same, the way in which the data is handled is completely different.

This article describes how natural language processing (NLP) technologies can be integrated in terminology work, especially in specialized dictionary compilation,[2] and shows how they can help terminologists extract or locate the data required for this kind of work.

[1] Observatoire de linguistique Sens-Texte (OLST), Université de Montréal, C.P. 6128, succ. Centre-ville, Montréal (Québec), H3C 3J7. mc.lhomme@umontreal.ca

[2] Other terminology projects involve the design and/or population of ontologies. I will not mention these projects as such, but most of what is said herein also applies to ontology projects.

2. Tasks in Specialized Dictionary Compilation

In order to compile dictionaries in special subject fields, terminologists must carry out a number of tasks. These can differ slightly if the dictionary is monolingual or bilingual. In monolingual projects, the main tasks can be divided as follows:

1. Build a specialized corpus;
2. Find relevant linguistic units, *i.e.* terms;
3. Handle term variants;
4. Find relevant contexts;
5. Locate semantically-related terms throughout the corpus;
6. Encode data in a (sometimes highly) structured environment.

In bilingual terminology projects, tasks 1 and 2 can differ in the sense that terminologists might want to build bilingual corpora (aligned or comparable) and align equivalents from these corpora rather than extract lists of terms independently in each language. These differences will be mentioned where applicable.

3. Automating the Compilation of Specialized Dictionaries

This section examines how the first five tasks listed in Section 2 can be automated. I will focus on these since they can be partly automated with NLP technologies. Examples are taken from two separate projects currently carried out at the Observatoire de linguistique Sens-Texte (OLST): 1) a terminological database on computing and the Internet; 2) a terminological database on climate change.

3.1. Build specialized corpora

Each time a new terminology project starts, this usually implies that a specialized corpus be built. In addition, the quality of results of other subsequent tasks depends heavily on the contents of the corpus. Hence, terminologists spend a considerable amount of time designing corpora for specific projects.

In order to help them locate relevant texts (*i.e.* texts that are likely to be useful for terminology work), a number of criteria have been defined and are listed in the literature (*cf.* Bowker and Person 2002, among others). These criteria include (but are not limited to): texts should be written by experts; texts should address specific topics (rather than general ones); texts should be up to date; texts should come from various sources. The ideal size of a specialized corpus remains an unsolved issue among terminologists, but it is a fact that specialized corpora are much most reduced in size than general reference corpora (used for lexicography work).

For the past decade or so, the challenge has been to apply these criteria to texts in electronic form. A first basic strategy would consist in browsing the Web using key words (or seed terms) that are representative of the subject field targeted by a specific

terminology project and enter them in a search engine. For example, for a dictionary project on climate change, seed terms such as *climate change* and *global warming* could serve as a basis to extract texts. By applying this simple technique, terminologists can locate relevant documents, but it is also likely that this will generate a lot of noise.

To find relevant documents and reduce the amount of noise, Agbago and Barrière (2005) developed a technique based on an analysis of some of the characteristics of documents found on the web.[3] The main feature the system looks for is the presence of linguistic markers (also called *knowledge patterns*, Meyer 2001, *cf.* Subsection 3.4) that indicate metalinguistic information (definitions, such as *is defined as*, *refers to*) or semantic relations (hyperonymy, such as *is a type of*; meronymy, such as *is a part of*). The method by Agbago and Barrière is based on the assumption that documents that contain targeted linguistic markers are likely to be more useful for terminology work than documents that do not. Terminologists enter seed terms, then the system automatically applies the list of markers, looks for documents that contain instances of terms and markers and ranks them according to the density of markers.

In bilingual projects, terminologists can build two different kinds of bilingual corpora, namely comparable corpora (that contain texts originally written in each of the languages described, but addressing the same topics and complying with the same criteria) or translated and aligned corpora (that contain source texts in one language and translated texts in the other). The use of translated corpora remains a controversial issue in some terminology circles. According to some authors, translated texts are not as authentic as non-translated texts since they were not written by experts in a given subject field.

3.2. Find relevant linguistic units

Once the corpus has been built and deemed satisfactory, the next step consists in identifying relevant linguistic units contained in the documents. These units, called *terms*, are likely to be those that will appear as headwords in dictionary articles.

This task is extremely difficult to automate. First, the definition of "term" can vary from one project to the other, since the notion of "term" depends heavily – although not exclusively – on a specific application. Secondly, an NLP technique will need to distinguish, among all linguistic units that appear in running text, those that comply with this specific definition of "term" from those that do not. The short extract reproduced in Figure 1 will serve to illustrate the complexity of this task. Most terminologists and experts would probably agree that units appearing in bold characters correspond to terms and should be identified at this stage of any terminology project. However, other units will inevitably lead to discussion (*cf.* units appearing in bold italics in Figure 1). Typical questions to be raised at this point are:

[3] The technique was implemented in a system called TerminoWeb that includes other functionalities designed to assist terminologists (*i.e.* term extraction and search for useful contexts).

Should we consider multi-word units only (*e.g. global warming*) or also include single-word terms (*e.g. warming*)? Should we consider nouns only (*e.g. global warming*, *warming*, *global cooling*) or also include other parts of speech (*e.g. to warm*, *to cool*).

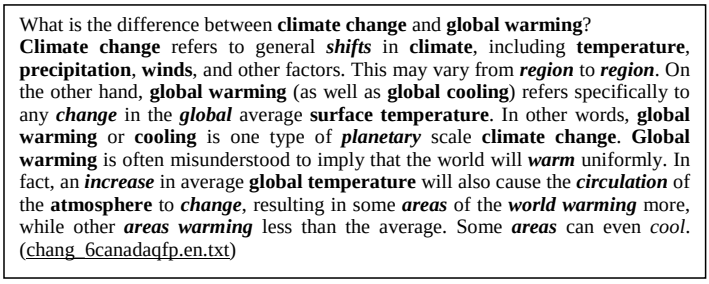

What is the difference between **climate change** and **global warming**?
Climate change refers to general ***shifts*** in **climate**, including **temperature**, **precipitation**, **winds**, and other factors. This may vary from ***region*** to ***region***. On the other hand, **global warming** (as well as **global cooling**) refers specifically to any ***change*** in the ***global*** average **surface temperature**. In other words, **global warming** or **cooling** is one type of ***planetary*** scale **climate change**. **Global warming** is often misunderstood to imply that the world will ***warm*** uniformly. In fact, an ***increase*** in average **global temperature** will also cause the ***circulation*** of the **atmosphere** to ***change***, resulting in some ***areas*** of the ***world warming*** more, while other ***areas warming*** less than the average. Some ***areas*** can even *cool*. (chang_6canadaqfp.en.txt)

Figure 1. Terms in a short extract on climate change

Since the beginning of the 90s, several techniques have been developed to automate, albeit partly, the identification of terms. Automatic term extraction produces lists of units, called *candidate terms* that are then analyzed by terminologists who will select those candidates that are valid (*i.e.* that are useful for a specific project).

I will present a single example of a technique that has been used at the OLST. The technique consists in comparing the frequencies of units that appear in different corpora. The technique itself was not designed for terminology applications as such, but Drouin (2003) applied it to term extraction and implemented it in a tool called *TermoStat*. The assumption on which the technique is based is that units that are proportionately more frequent in a specialized corpus (*e.g.* a corpus on computing) than in a reference corpus (*e.g.* a newspaper article corpus) are likely to be terms. The technique can be applied to single-word terms as well as multi-word terms and to all parts of speech. Appendices 1 and 2 show some results that we obtained when comparing corpora in English and in French.

In bilingual terminology projects, terminologists can try to identify potential equivalents of candidate terms right away instead of extracting terms from each corpus independently. They can be assisted in this task by lexical or terminological aligners. Aligning candidates combine the problems of term extraction (distinguishing terms from other linguistic units in running text) with the problems of lexical alignment *per se* (finding equivalent words in sentences in which the order of words is not the same, for example), which is not a trivial task. Equivalent terms themselves may display differences that will further complicate the task of recognizing them automatically: different structures (En. *computer-assisted terminology*: Fr. *terminotique*; En. *to bookmark*; Fr. *mettre en signet*); different parts of speech (Fr. *climatique*: En. *climate*; Fr. *informatique* (adj.); En. *computer*); more than one equivalent in a corpus (Fr. *anthropique*: En. *anthropogenic*, *human-induced*).

One possible way to automate this task consists in combining two tools: a term extractor (TermoStat, Drouin 2003) and a lexical aligner (Alinea, Kraif *et al.* 2004) that was not designed for term alignment only. We carried out an experiment (Le Serrec *et al.* 2010, forthcoming) that shows the benefits of this combination. Appendix 3 illustrates the ordering of terms when extracted by TermoStat separately from an English corpus on climate change and the corresponding French corpus. Appendix 4, on the other hand, shows the results of the automatic alignment performed by Alinea when fed with the French candidate terms. Many ordering problems are corrected and more than one possible equivalent was found by the aligner for some French candidates.

3.3. Handle term variants

When identifying terms in running text, and especially when trying to extract them automatically, terminologists inevitably run into the problem of handling term variants, *i.e.* different linguistic expressions that refer to the same concept or that convey the same meaning. Various typologies of variants are available in the literature, and I reproduce one of them below (each type of variant will be handled differently by an automatic process and some are obviously more complex to handle then others):

- Graphical variants: *blogueur*, *blogger*, *bloggueur*; *carbon dioxyde*, *CO2*
- Inflected forms: *change* vs. *changes*; *warm*, *warmed*, *warming*
- Morphological and morphosyntactic variants: *décrypter* vs. *crypter*; *climate change* vs. *the climate is changing*; *climate change* vs. *climatic change*
- "Light" syntactic variants and syntactic variants per se: *imprimante à laser* vs. *imprimante laser*, *global warming or cooling = global warming, global cooling*
- Synonyms: *area* vs. *region*; *courriel* vs. *mail*

From the point of view of term extraction, variants raise two different problems:

1. Variants can appear in very different parts of a list of candidate terms and terminologists will need to group them manually. For example, in a list of candidate terms sorted by descending order of frequency, the variants of *climate change* and *warm* will appear as follows (imagining that other candidates will appear between them as well):

   climate change (50)
   …
   warm (12)
   …
   warms (6),
   change in climate (5)
   warmed (5)
   …
   climatic change (3)
   warming (2)

2. Even worse still, depending on the rules defined in the term extractor, some occurrences of terms might be missed. For example, in the following coordination, *climate and weather changes*, term extractors might be able to catch *weather change*, but not *climate change*, since the head *change* appears only once.

Different methods exist for handling specific kinds of term variants. Graphical variants and inflected forms are now processed with standardization and lemmatization in most term extractors. The addition of a good lemmatization method to a term extractor gives a much more accurate picture of the frequency of terms instead of giving the separate frequencies of each of their inflected forms. For example, a list of term candidates will state that the verb *warm* appears 15 times in a corpus instead of stating that *warmed* has a frequency of 5, *warming*, a frequency of 2, and *warms* a frequency of 2, etc.

Morphological variants must be handled with more sophisticated methods. Daille (2001) developed a method that groups terms that share the same head but in which the modifier differs: in the first term, the modifier is a noun (*e.g.*, *climate change*); in the second one, the modifier is a relational adjective derived from this noun (*e.g.*, *climatic change*).[4] The method identifies potential relational adjectives in extracted multi-word terms based on morphological rules and then finds terms that contain nouns from which these adjectives are derived. If a match is made, then terms are grouped and presented as variants to the user.

Syntactic variants can be handled by sets of rules that allow components of multi-word terms to be inverted, split, etc. Efficient methods were devised by Jacquemin (2001) and Ville-Ometz *et al.* (2007), among others.

3.4. Find relevant contexts

Much of the information that can help terminologists better understand the meaning of selected terms or identify relationships between concepts in the subject field is found in the corpus. Once the list of terms is defined – even if provisionally – terminologists go back to the corpus with the objective of finding contexts that contain this kind of information. Looking for relevant contexts in large corpora can be a very time-consuming task, since a basic concordancer will inevitably extract all sorts of sentences, some of which will be useful, but many of which will not contain information for which terminologists are looking.

Terminologists are confronted with two basic problems. First, useful contexts (for terminology) must be distinguished from others that terminologists will discard. For example, only the first of the two following sentences is interesting for a terminologist who seeks information about "climate change".

[4] The method was originally developed for French in which this transformation is very frequent in specialized corpora.

> Climate change refers to general shifts in climate, including temperature, precipitation, winds, and other factors.
>
> The sixth environment action programme identifies four priorities: climate change, nature and biodiversity, environment and health, and quality of life.

Secondly, terminologists might want to locate contexts that express a specific kind of information while ignoring others. For example, the first context below is interesting in the sense that it expresses a cause-effect relation; the second is also interesting since it gives a potential hyperonym for *climate change*; the third one would not be considered as a useful context in terminology work.

> Climate change causes extreme weather events
>
> Climate change is one of the greatest environmental, social and economic threats facing the planet. During the 20th century
>
> Climate change is a complex issue

One well documented strategy to locate useful contexts consists in exploiting linguistic markers that appear in metalinguistic contexts or that express a specific relation. Over the years, lists of markers have been compiled for different languages. Below are a few examples of such markers for English based on lists provided in Hearst (1992), Ahmad *et al.* (1994), Barrière (2004), Meyer (2001), Auger and Barrière (2008), etc.

- Markers for definitional contexts: *X is defined as Y, Y is called Y*
- Markers for hyperonymy: *X is a type of Y, X and other Ys*
- Markers for meronymy: *X is a part of Y, Y consists of X …*
- Markers for cause-effect: *X causes Y, X impacts Y…*
- Markers for antonymy: *X differs from Y, X is different from Y*

These markers can be used in two different ways. First, terminologists can enter a list of markers (say for hyperonymy) along with a given term in a concordancer and have the concordancer extract those contexts that contain the term and an occurrence of a marker. Secondly, markers can be implemented in a system designed for terminology work (as was done in TerminoWeb, Barrière 2004). Users enter a term or a list of terms and a search for these terms in contexts in which markers appear is performed automatically. Below is an example of the kinds of results that could be obtained using a character string approach. In this example, contexts for *microprocessor* are sought. The first two sentences express meronymic relations and the next two hyperonymic relations. The last two contain *microprocessor* as well as markers but the relation does not concern the term itself: these latter errors can be avoided if a syntactic analysis is superimposed to the character recognition.

> In some embodiments of the system the **microprocessor is part of** a microcontroller
>
> If the **microprocessor is part of** an irrigation controller, the microprocessor
>
> A **microprocessor is a type of** integrated circuit or chip and is the heart of every computer.

A **microprocessor is a type of** electronic device that provides intelligence
Study of **Microprocessor is a part of** the most Engineering Curriculums.
CMOS, **is a type of** IC which includes **microprocessor**, microcontroller

3.5. Locate semantically related terms throughout the corpus

The methods described in the previous subsection assume that relationships between terms are expressed in the same vicinity (*i.e.* in the same sentence). However, many relationships exist between terms that can appear in very different parts of the corpus (in different sentences, in different paragraphs or even in different texts). This is the case for relationships between synonyms (*e.g. region* vs. *area; email* vs. *mail*), between antonyms or contrastive terms (*e.g. install* vs. *uninstall*; *analog* vs. *digital*), or other types of paradigmatic relationships (*e.g. changement climatique, refroidissement, réchauffement*).

To locate pairs of terms sharing these relationships, different methods were developed. Two examples are presented in this subsection. The first one consists in using a list of multi-word terms extracted from a corpus, locate those candidates that have the same components and reorganize the candidates on this basis. The method was implemented in a number of term extraction systems. It is based on the assumption that terms that share formal components also share semantic features. Bourigault (1994) presents a version of the method with an additional constraint: the components must share the same syntactic function. Figure 2 shows terms containing *change* (as a head or as a modifier) that were reorganized in order to better display their similarities. Of course, any other component can lead to another series if it appears in more than one multi-word term.

climate	change		
land-use	change		
temperature	change		action
	change		activity
	change		adaptation
climatic	change		
environmental	change		
global	change		
initial	change		
technical	change		
technological	change	in	composition
	change	of	climate

Figure 2. Terms sharing the same head or modifier

Another method, developed by Claveau and L'Homme (2005), consists in finding single-word terms that are morphologically related in a list of extracted term candidates. In addition to grouping them, the method suggests labels for the relationships. The method is divided into the following steps:

- Look for pairs of terms that appear in a pre-existing dictionary of computing: (the terms must contain a common character string): *activer – activable*.
- Extract the semantic relationship between the terms: $Able_2$ (lexical function[5] that means basically the adjective that means "that can be verb-ed").
- Infer rules for pairs: $Able_2$(w1) = w2 if w1 -suf “er” +suf “able” = w2 (this rules states that, in order to obtain an $Able_2$ relationship between word 1 and word 2 – that have a given morphological distance –, the suffix *–er* must be removed from word 1 and the suffix *–able* must be added to word 1).
- Apply these rules to a list of extracted terms: the rules are applied using a technique called *analogy* (Lepage 2003; A is to B what C is to D): the rule inferred for the relationships between *activer* and *activable* is valid for other pairs of terms, namely, *compiler* and *compilable.*

Figure 3 shows some results that can be obtained using this technique for terms belonging to the *compil-* family. The analysis was carried out for French terms. The English equivalents are given for information purposes. In these results, the only invalid links are those found between *compiler -> compilation* and *recompiler -> recompilation*).

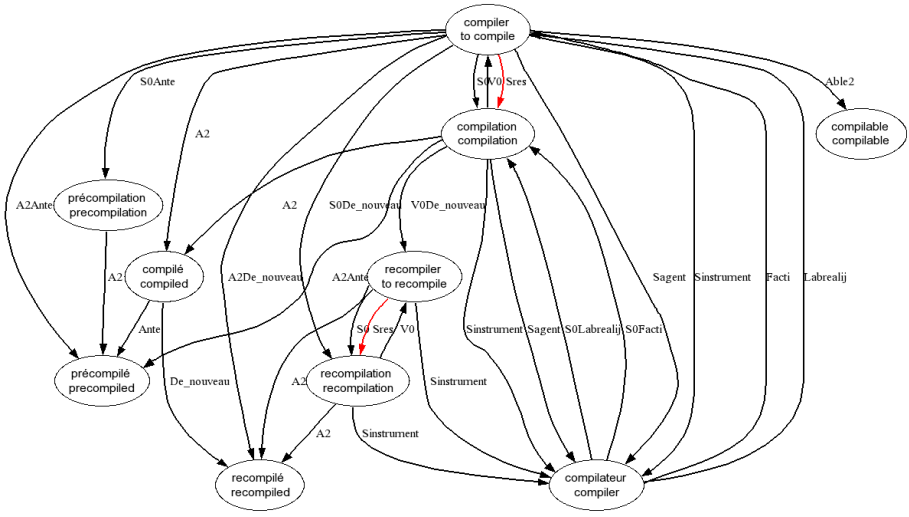


Figure 3. Morphological relationships between terms pertaining to the compil- *family* (Claveau and L’Homme 2005)

[5] In the dictionary, semantic relationships are represented using the system of lexical functions (Mel’čuk *et al.* 1995).

4. Concluding remarks

In this article, I gave a very brief account of some NLP methods that are or that can be used in terminology work to alleviate some of the tasks terminologists must carry out. The combined use of these methods gives terminologists different views on the data contained in specialized corpora.

Methods were presented for what they are and I deliberately avoided addressing one very important question for practitioners who use them in everyday work: do these technologies really improve terminology work? Since none of the techniques presented above produces perfect results, some question their usefulness.

Advocates of the incorporation of NLP methods in terminology work argue that they are extremely useful to process large corpora in a much more systematic way. They can also help terminologists handle different sets of data separately instead of having to browse running text and having to make decisions about many different problems at the same time. Other practitioners wonder whether NLP techniques can really help them work faster. For example, considering that a typical term extractor generates a lot of noise (invalid candidates) and misses some useful terms, a lot of time must be spent on browsing the list and removing erroneous candidates. In addition, missed terms must be extracted manually.[6]

Of course, many other NLP techniques currently developed can be applied to terminology work. Here is a list of some current developments than could help alleviate other tasks in terminology: extension of equivalent alignment to comparable corpora instead of being limited to aligned corpora; extension of variant recognition and semantic relationship identification to bilingual corpora; collocation extraction; automatic identification of the argument structure of terms along with the labelling of the semantic roles of arguments; building of semantic classes of terms based on distributional analysis, etc.

Even if practitioners disagree on the usefulness of NLP technologies in general or in relation to some specific tasks, all accept the fact that NLP and resources in electronic form have completely changed the way data in corpora is analyzed. They also made terminologists consider types of data they overlooked before. The tendency now is not to question the presence of NLP tools (it is now taken for granted), but rather how terminologists can interact with them in order to facilitate some of their tasks.

References

AGBAGO, A. and BARRIÈRE, C. (2005). Corpus Construction for Terminology. In *Corpus Linguistics 2005 Conference*, Birmingham, UK.

[6] Due to space limitations, I cannot mention the evaluations that were carried on the applications I presented in these pages: a complete evaluation of TerminoWeb is reported in Barrière (forthcoming); TermoStat was evaluated in Drouin (2003) and Lemay *et al.* (2005); the combined use of TermoStat and Alinea was evaluated and will be reported in Le Serrec *et al*. (2010, forthcoming).

AHMAD, K. and FULFORD, H. (1992). *Knowledge Processing: 4. Semantic Relations and their Use in Elaborating Terminology*. (Computing Sciences Report CS-92-07). Guildford.

AUGER, A. and BARRIÈRE, C. (eds). (2008). *Pattern-based Approaches to Semantic Relation Extraction*. Special issue of *Terminology*, 14(1).

BARRIÈRE, C. (2004) Knowledge-Rich Contexts Discovery, In *Proceedings of the 17th Canadian Conference on Artificial Intelligence (AI'2004)*, London, Ontario.

BARRIÈRE, C. (forthcoming). *TerminoWeb 1.0. Évaluation et Perspectives*.

BOURIGAULT, D. (1994). *LEXTER. Un logiciel d'EXtraction de TERminologie. Application à l'acquisition de connaissances à partir de textes*. Thèse de doctorat. Paris: École des hautes études en sciences sociales.

BOWKER, L. and PEARSON, J. (2002). *Working with Specialized Language. A Practical Guide to Using Corpora*. London: Routledge.

CLAVEAU, V. and L'HOMME, M.-Cl. (2005). Terminology by Analogy-Based Machine Learning. In *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering*, TKE 2005, Copenhagen.

DAILLE, B. (2001). Qualitative term extraction. In D. Bourigault., C. Jacquemin and M.-Cl. L'Homme (eds). *Recent Advances in Computational Terminology*. Amsterdam / Philadelphia: John Benjamins: 149-166.

DROUIN, P. (2003). Term Extraction using Non-technical Corpora as a Point of Leverage. *Terminology*, 9(1): 99-115.

HEARST, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING-92*. Nantes: 539-545.

JACQUEMIN, C. (2001). *Spotting and Discovering Terms through NLP*. Cambridge: Cambridge University Press.

KRAIF, O. and CHEN. B. (2004). Combining clues for lexical level aligning using the Null hypothesis approach. In *Proceedings of Coling 2004*. Geneva: 1261-1264.

L'HOMME, M.-Cl. (2004). *La terminologie: principes et techniques*. Montréal: Presses de l'Université de Montréal.

LE SERREC, A., L'HOMME, M.-Cl., DROUIN, P. and KRAIF, O. (2010, forthcoming). Automating the compilation of specialized dictionaries: use and analysis of term extraction and lexical alignment. *Terminology* 16(1).

LEMAY, C., L'HOMME, M.-Cl. and DROUIN, P. (2005). Two Methods for Extracting "Specific" Single-word Terms from Specialized Corpora: Experimentation and Evaluation. *International Journal of Corpus Linguistics*, 10(2): 227-255.

LEPAGE, Y. (2003). *De l'analogie; rendant compte de la communication en linguistique*. Grenoble.

MEL'CUK, I., CLAS, A. and POLGUERE, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve: Duculot / Aupelf - UREF.

MEYER, I. (2001). Extracting knowledge-rich contexts for terminography. In D. Bourigault., C. Jacquemin and M.-Cl. L'Homme (eds). *Recent Advances in Computational Terminology*. Amsterdam/Philadelphia: John Benjamins: 279-302.

VILLE-OMETZ, F., ROYAUTÉ, J. and ZASADZINSKI, A. (2007). Enhancing in automatic recognition and extraction of term variants with linguistic features. *Terminology*, 13(1): 35-59.

Appendix 1: First candidate terms obtained when comparing an English specialized corpus on climate change and the British National Corpus.

Single-word terms and nouns, verbs, adjectives and adverbs were considered. Erroneous or questionable candidates are in bold.

TOKEN	PART OF SPEECH	FREQUENCY	WEIGHT
Climate	Noun	520	235,082665840696
Change	Noun	4256	188,691871033502
Emission	Noun	3049	182,802936994759
Global	Adjective	1667	128,652217650742
Temperature	Noun	1733	126,266619096255
Model	Noun	1669	117,573341384666
Scenario	Noun	1228	114,448844906524
Carbone	Noun	1268	114,077796151994
Greenhouse	Noun	1299	113,25659323117
Gas	Noun	1654	111,832214129543
%	**Noun**	**1173**	**111,320882199399**
Concentration	Noun	1088	102,596981373238
Ocean	Noun	1041	101,06077145499
Impact	**Noun**	**1215**	**98,6294939933866**
Atmosphere	Noun	1017	94,9008316617182
Warming	Noun	853	94,1821405406052

Appendix 2: First candidate terms obtained when comparing a French specialized corpus on computing and the newspaper Le Monde.

Single-word terms and all parts of speech were considered. Erroneous or questionable candidates are in bold.

TOKEN	PART OF SPEECH	FREQUENCY	WEIGHT
Fichier	SBC	3956	360.825
Commande	SBC	1902	201.749
Option	SBC	1486	182.338
Serveur	SBC	1166	180.477
Utilisateur	SBC	1117	167.307
Configuration	SBC	845	162.82
Utiliser	**VB**	**1996**	**161.788**
Répertoire	SBC	1003	153.668
Système	SBC	2699	152,979
Disquette	SBC	609	148.431
Windows	SBP	613	124.412
Votre	**DT**	**1365**	**124.361**
Disque	SBC	1096	120.167
Réseau	SBC	1511	119.167
Imprimante	SBC	537	117.294
Mémoire	SBC	1240	112.42

Appendix 3: English and French candidate terms extracted from two corpora

Ordering problems start at the fourth position.

French candidate	Rank in the list	Rank in the list	English candidate
Climatique	1	1	Climate
Changement	2	2	Change
Émission	3	3	Emission
Température	*4*	*4*	*Global*
Carbone	*5*	*5*	*Temperature*
Climat	*6*	*6*	*Model*
Serre	*7*	*7*	*Scenario*
Gaz	*8*	*8*	*Carbon*
Réchauffement	*9*	*9*	*Greenhouse*
Forçage	*10*	*10*	*Gas*
Co2	*11*	*11*	*%*
Atmosphère	*12*	*12*	*Concentration*

Appendix 4: French candidate terms aligned automatically with English equivalents with Alinea

French candidate	Rank in the list	Rank in the list	English candidate
Climatique	1	1 52	Climate climatic
Changement	2	2	Change
Émission	3	3	emission
Température	4	5	temperature
Carbone	5	8 27/130	Carbon co2
Climat	6	1	climate
Serre	7	6 413/438	Greenhouse GHG/GHGs
Gaz	8	10 413/438	Gas GHG/GHGs
Réchauffement	9	16 161 1745	Warming Warm (to) warmer

Word frequency distribution for electronic learner's dictionaries

Hanhong Li[1]
City University of Hong Kong

Abstract

Word frequency information has been an indispensable part of electronic dictionaries for learners of English. A review of the current five major electronic learner's dictionaries (*LDOCE5*, *OALD7*, *CALD3*, *MED2*, *COBUILD5*) shows that word frequency information is mainly based on raw frequency without considering distributed frequency across different genres. Our research explores the distributed frequency in 14 reorganized genres (9 written, 5 spoken) of the British National Corpus XML Edition (BNC XML 2007). Statistics show that a core vocabulary for EFL learners selected by distributed frequency achieves higher cumulative coverage than a core vocabulary selected by raw frequency alone. An electronic *English Frequency Dictionary* has been developed to display distributed frequency across genres as a histogram. This software helps EFL learners to understand the register and cultural implication of English words, and to identify some subtle distinctions between synonyms. This enhancement is proposed for future electronic learner's dictionaries.

Keywords: distributed frequency, genre, core vocabulary, BNC, frequency dictionary.

1. Word frequency in electronic learner's dictionaries

Word frequency information has been indispensable in building electronic English learner's dictionaries, and has been used to select headwords, order senses, identify collocations, and even define core vocabulary for definitions. In the current five major electronic learner's dictionaries (*LDOCE5*, *OALD7*, *CALD3*, *MED2*, *COBUILD5*) word frequency information is mainly based on raw frequency (or total frequency) without considering the distributed frequency in different genres. Although corpora have been widely used for dictionary compilation, Bogaards (2008) pointed out that the current learner's dictionaries mainly focus on written data without revealing the proportion of the spoken language used in their corpora. His comparison between the data presented in the five major learner's dictionaries even cast some doubt on the reliability of their frequency indication. Two approaches to frequency marking will be summed up here after the review of the *Longman Dictionary of Contemporary English 5th Edition* (*LDOCE5*), the *Cambridge Advanced Learner's Dictionary 3rd Edition* (*CALD3*), the *Macmillan English Dictionary 2nd Edition* (*MED2*), the *Collins*

[1] Department of Chinese, Translation and Linguistics, johnlihanhong@yahoo.com.cn

COBUILD Dictionary on CD-ROM 2006 (*COBUILD5*) and the *Oxford Advanced Learner's Dictionary 7th Edition* (*OALD7*).

1.1. Raw frequency

The frequency marking in current dictionaries is based only on the counts of tokens in large corpora, viz. raw frequency. As summed up in Table 1, *CALD3* presents three levels of frequency: 1) essential level, which indicates a common, useful and important word to know; 2) improver level, which refers to words to help users improve beyond basic English; 3) advanced level, which describes those words that make your English sound advanced. *MED2* designates three groups of word frequency: very high, high and quite high according to their commonness. *COBUILD5* shows three word frequency bands with diamond symbols. The most frequent words have three diamonds, the next most frequent words have two diamonds, and those with lower frequency have one diamond.

Electronic Learner's Dictionary	**Frequency Symbol**	**Frequency Indication**
LDOCE5	W1 W2 W3 S1 S2 S3	W1: Most frequent 1000 written words W2: Between 1001 and 2000 most frequent written words W3: Between 2001 and 3000 most frequent written words S1: Most frequent 1000 spoken words S2: Between 1001 and 2000 most frequent spoken words S3: Between 2001 and 3000 most frequent spoken words
OALD7	🔑	Words in Oxford 3000 which are frequent across a range of different types of texts
CALD3	❶ ❶ ❶	Essential: a common, useful and important word to know Improver: a word to help you improve beyond basic English Advanced: a word to make your English sound advanced
MED2	★★★ ★★ ★	3-star: the 2500 most common and basic English words 2-star: very common words 1-star: fairly common words
CCD5	◆◆◆ ◆◆ ◆	The most frequent words have three diamonds, the next most frequent two, and the ones which are less frequent have one diamond

Table 1. Frequency marking of headwords in current electronic learner's dictionaries

1.2. Frequency in range level

Besides the primitive raw frequency information, some learner's dictionaries have employed frequency with range information. In frequency counting, range refers to the number of text categories or samples in which a word occurs (Juilland and Chang-Rodriguez 1964). Thorndike and Lorge (1944) considered the range credit in the

vocabulary control movement. Kučera and Francis (1967) and Hofland and Johansson (1982) introduced range statistics of different text categories in their wordlists. Range has been an influential factor for frequency evaluation. Kilgarriff (1997) demonstrated that the marking of range statistics in dictionaries could present more information than those without this data. Based on the range of spoken and written English, *LDOCE5* progressed to mark the 3000 most frequent spoken and 3000 most frequent written words (named the Longman Communication 3000). Words in the Longman Communication 3000 have to be widely distributed across a wide range of sources (Bullon and Leech 2007). *OALD7* also considered the range factor when it selected words for its core word list titled the Oxford 3000. This core vocabulary list only includes those words which are frequent across a range of different text types.

1.3. Frequency for defining vocabulary

Frequency information is not only found in the frequency marking of headwords in these five learner's dictionaries but also in the design of their defining vocabulary. The concept of defining vocabulary was first introduced by West (1935) and made its debut in *The New Method English Dictionary* (Cowie 1999). *LDOCE5*, *OALD7* and *MED2* have developed their own defining vocabularies. Words in the defining vocabulary of *LDOCE5* "are made up of very frequent words" (Bullon and Leech 2007: 6) and are "constantly being researched and checked to make sure that they are frequent in the Longman Corpus Network" (26). The defining vocabulary of *MED2* contains the most common and basic words in English and they have been chosen by examining the word frequency information from hundreds of millions of words of English (*MED2*). The defining vocabulary in *OALD7* is named the Oxford 3000 and only includes those words which are frequent across a range of different text types. Generally speaking, the major selecting criteria of defining vocabularies are divided into two categories: 1) raw frequency (frequent in large corpora) *e.g. LDOCE5* and *MDE2*; and 2) distributed frequency (frequent across a range of text types) *e.g. OALD7*.

1.4. Frequency problem

Although distributed frequency in different text types has been adopted by *OALD7* to establish its core vocabulary, the Oxford 3000, no distributed frequency has been indicated for headwords in electronic learner's dictionaries. Moreover, the current electronic learner's dictionaries never explicitly state the following details: What is the definition of their text types? How many and what text types have they used? How many types of texts should a word occur in before it is selected for frequency marking? How large are the sub-corpora for each text type? Is the variety of written and spoken texts enough? Are they balanced? Without this kind of detailed information, some scholars (*e.g.* Bogaards 2008) have started to challenge the reliability of the frequency bands.

Genres in BNC	Super-genres	Token Size
W_ac_humanities_arts (academic prose: humanities)	w_academic	1012585
W_ac_medicine (academic prose: medicine)		
W_ac_nat_ science (academic prose: natural sciences)		
W_ac_soc_science (academic prose: social & behavioural sciences)		
W_ac_tech_engin (academic prose: technology, computing, engineering)		
W_ac_polit_law_edu (academic prose: politics, laws, education)		
W_non_ac_humanities_arts (non-academic/non-fiction: humanities)	w_nonAcademic	1050676
W_non_ac_medicine (non-academic: medical/health matters)		
W_non_ac_nat_science (non-academic: natural sciences)		
W_non_ac_soc_ science (non-academic: social & behavioural sciences)		
W_non_ac_tech_engin (non-academic: technology, computing, engineering)		
W_non_ac_polit_law_edu (non-academic: politics, law, education)		
W_religion (religious texts, excluding philosophy)	w_religion	1014956
W_pop_lore (popular magazines)	w_pop	1024675
W_fict_prose (novels & short stories)	w_fict_pros	1012716
W_newsp_brdsht_nat_report (broadsheet national newspapers: home & foreign news reportage)	w_newsp_report	1012637
W_newsp_other_report (regional and local newspapers: home & foreign news reportage)		
W_biography (biographies/autobiographies)	w_biography	1004618
W_commerce (commerce & finance, economics)	w_commerce	1021022
W_admin (administrative and regulatory texts, in-house use)	w_public	1001948
W_hansard (Hansard/parliamentary proceedings)		
W_institut_doc (official/govermental documents/leaflets, company annual reports, etc.; excludes Hansard)		
S_demg_daily conversation (face-to-face spontaneous conversations)	s_conversation	1039168
S_cg_Educational/Informative	s_education	1001780
S_cg_Business	s_business	1004478
S_cg_Public/Institutional (political, public, religion)	s_public	1022673
S_cg_Leisure	s_leisure	1000289
TOTAL		14224221

Table 2. Corpus of balanced super-genres

2. Frequency Distribution in Various Text Genres

What kind of frequency information will be more helpful for learners: raw frequency or distributed frequency in different text types? With the further development of modern corpora and studies in text genres, a new vision of word frequency information is available for our dictionary users: word frequency distribution in different genres.

2.1. Purpose

This study explores the application of distributed frequency in various genres of a large corpus. We aim to 1) explore the strengths of distributed frequency when selecting defining vocabulary or core vocabulary; 2) compile an electronic dictionary with frequency distributed in different text genres which could provide more helpful information than current electronic learner's dictionaries, with a view to advocating this kind of distributed frequency technique for future learner's dictionaries.

2.2. Genre Corpus

In order to obtain the distributed frequency in large corpora, we have to set up a corpus containing various text genres. The British National Corpus XML (BNC XML 2007) comprises 100 million words. The spoken component of the BNC contains approximately 10 million words and the written component 90 million words. This corpus has been tagged with 70 genres, *i.e.* 24 spoken genres and 46 written genres (Lee 2001). By reorganizing and combining some of the current genres, we created 9 written super-genres (academic prose, non-academic prose, biography, fiction, news, public, religion, popular magazines, and commerce) and 5 spoken super-genres (conversation, education, business, public, and leisure). Then we selected around 1 million word units for each of the aforementioned super-genres and established a balanced genre corpus of 14 genres to explore the word frequency distribution across them (see Table 2).

2.3. A statistical method for distributed word frequency in different genres

In order to calculate the distributed frequency in different genres, we adopt Carroll's (1970) statistical method. Carroll's usage coefficient, *Um*, in the following formula is used to calculate the distributed frequency value in different text categories of a corpus:

F = the total frequency of the given word in the corpus

N = the total number of tokens in the corpus

n = the number of categories

f_j = the sub-frequency of a given word-type in category j (j~ 1, 2, ..., n)

S_j = the number of tokens in category j

N = $\sum S_j$ = the total number of tokens in the corpus

$p_j = f_j / S_j$ = the proportion of tokens in category *j* that are instances of the given word-type

$P = \sum_j p_j$ (*P* may take any positive value; it is not in general equal to unity.)

Then:

$H = \log P - \frac{\sum_j p_j \log p_j}{P}$ ($p_j \log p_j = 0$ for $p_j = 0$)

$D_2 = H/\log n$ = the index of dispersion

$f_{min} = (\sum S_j f_j) / N$

$Um = (1{,}000{,}000/N)\ [FD_2 + (1\text{-}D_2) f_{min}]$

With the above formula, we can calculate the *Um* value for each word-type in the genre corpus established in 2.2. *Um* is the usage coefficient and can be used to indicate the frequency value which considers the frequency distributed in different text genres.

2.4. Comparison of core vocabulary lists

In order to evaluate the different effects between raw frequency and distributed frequency, we compared the cumulative coverage between two core vocabulary lists which selected words by two different criteria: raw frequency *vs.* distributed frequency.

Since defining vocabulary has been regarded as a kind of core vocabulary in previous research (Lee 2001), we adopted the Longman Defining Vocabulary (LDV) in *LDOCE5* as a core vocabulary. As stated in Section 1.3, the word selection criteria of the LDV mainly depend on raw frequency. The LDV also includes 30 affixes which can combine with the existing words in the LDV to create more possible words for its defining purpose. The possible derivatives in the LDV should be considered so as to evaluate the power of the LDV (Herbst 1986). We used the advanced search function in the electronic *LDOCE5* to search for all the used derivatives in the Longman definitions. Phrases were not considered at this stage. Finally, we obtained 2,751 words from the LDV (see Table 3).

Core Vocabulary	**Word Number**	**Criteria**	**Cumulative Coverage in Genre Corpus (%)**
Longman Defining Vocabulary (LDV)	2751	Frequent (raw frequency)	83.55
Genre Core Vocabulary(GCV)	2751	Frequent in various genres (distributed frequency)	90.88

Table 3. Contrast of raw frequency and distributed frequency in core vocabulary

Then we created the other core vocabulary list named Genre Core Vocabulary (GCV) by selecting the same number of words as in the LDV from the word list of the genre corpus described in Section 2.2. First, we compiled a wordlist of the genre corpus and calculated the distributed frequency (*Um*) for each word with the formula given in Section 2.3. Second, we sorted the value of the distributed frequency (*Um*) in descending order and selected the first 2,751 words which represent the most frequent and best distributed words among the 14 genres in the corpus.

2.5. Data analysis

Table 3 shows that the core vocabulary list selected by distributed frequency produces better cumulative coverage than the core vocabulary list selected by raw frequency. Words with high frequency and even distribution in different text genres should be more reliably rated as core words than those which are only frequent in terms of raw frequency (*cf*. Carter 1998, Stubbs 2001).

3. Electronic frequency dictionary for learners

With the data in Section 2.3, we compiled an electronic *English Frequency Dictionary* to further explore the function of distributed frequency for learners.

3.1. Cultural implication and register

In translating the Chinese sentence 他是个乖孩子 (which literally means "he is a well-behaved child") into English, learners may turn to a Chinese-English dictionary. There they will find *good*, *well-behaved* or *obedient* for the translation of 乖. Without much thought, they may just select any of these translations. However, if they compare the respective frequency in Figure 1 and Figure 2 as produced by our electronic frequency dictionary, they can see that *good* is much more widely distributed in various genres than *obedient*. Furthermore, they will find that *obedient* is much more frequent in the genre of religion due to the influence of Christianity in Britain. However, this kind of difference is not mentioned in most electronic learner's dictionaries. Only *LDOCE5* mentions the difference as follows:

> **REGISTER**
> In everyday English, people usually say that a child is **good** rather than **obedient**:
> ▪ *The children were all very good.*

3.2. Synonyms

There are many synonyms in English. Is there any difference between them? This is what users really want to know. Figures 3 and 4 show the difference between the synonymous adjectives *naughty* and *mischievous*. Figure 3 tells us that *naughty* occurs more frequently in the genre of daily conversation rather than other context-governed spoken genres or any written genres. It is therefore informal. On the other hand, Figure

4 shows that *mischievous* occurs more frequently in written than spoken genres. However, this type of information is not included in any of the current five electronic learners' dictionaries. Only *OALD7* indicates that they are synonyms.

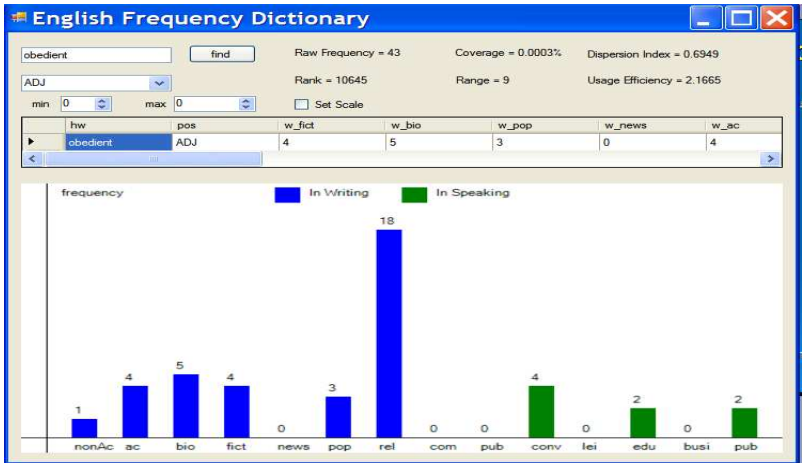


Figure 1. Frequency of obedient

nonAc: non-Academic prose ac: academic prose bio: biology fict: fiction
news: news report pop: popular magazines rel: religion com: commerce
pub: public conv: conversation lei: leisure edu: education busi: business

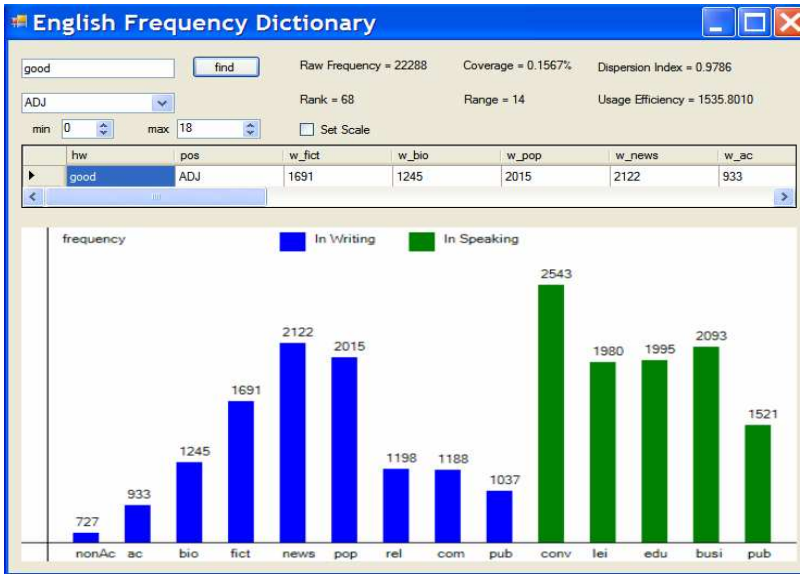


Figure 2. Frequency of good

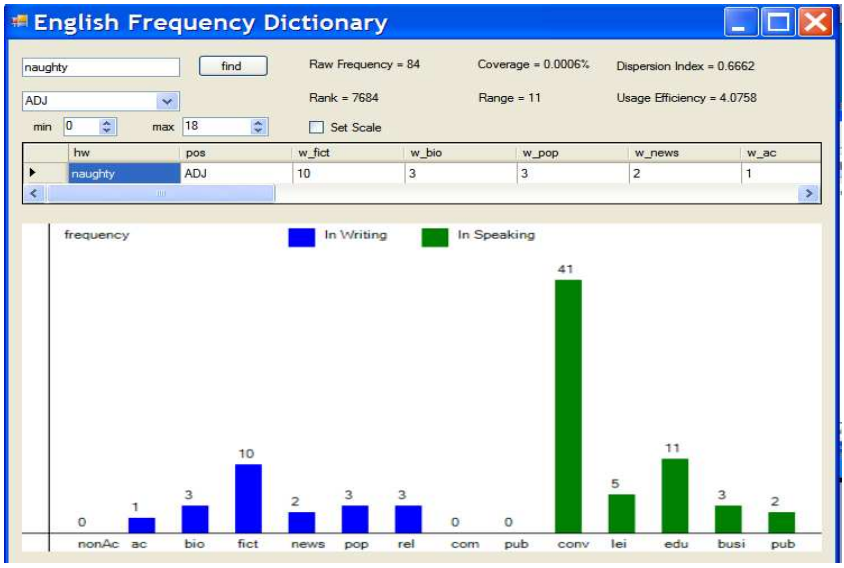


Figure 3. Frequency of naughty

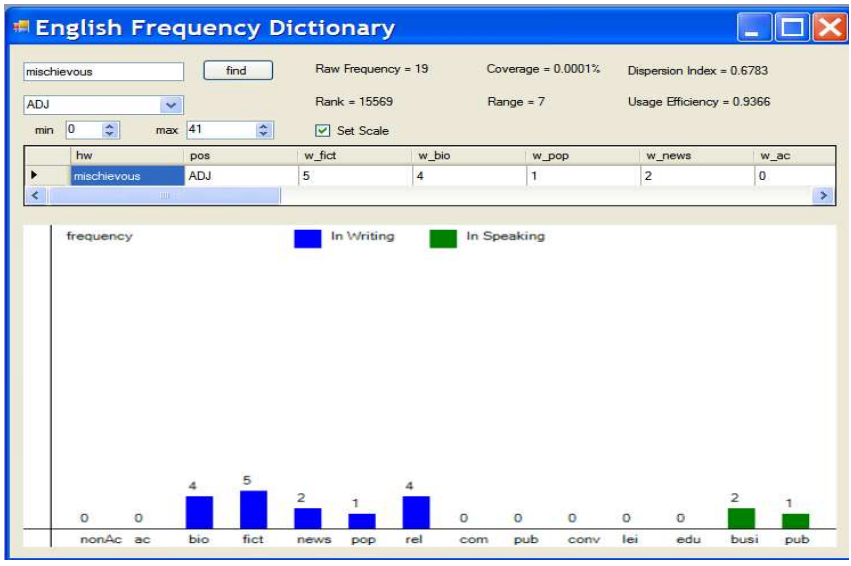


Figure 4. Frequency of mischievous

4. Conclusion

When we select words for defining vocabulary or core vocabulary, the criterion of distributed frequency in different text genres can achieve better coverage than raw frequency. Moreover, with interactive graphs of distributed frequency in a variety of text genres, English learners can infer and understand the register and cultural implications of English words, and even gain a better understanding of the differences between English synonyms.

References

BOGAARDS, P. (2008). Frequency in learners' dictionaries. In E. Bernal and J. DeCesaris (eds). *Proceedings of EURALEX-2008*. Barcelona: IULA, DOCUMENTA UNIVERSITARIA: 1231-1236.

BULLON, S. and LEECH, G. (2007). Longman Communication 3000 and the Longman Defining Vocabulary. In *Longman Communication 3000*. Harlow: Pearson Education Limited: 1-7.

CARTER, R. (1998). *Vocabulary: Applied Linguistics Perspectives, 2nd ed*. London and New York: Routledge.

CARROLL, J.B. (1970). An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a Standard Frequency Index (SFI). *Computer Studies in the Humanities and Verbal Behavior*, 3: 61-65.

COWIE, A.P. (1999). *English Dictionaries for Foreign Learners: a History*. Oxford: Oxford University Press.

HERBST, T. (1986). Defining with a controlled defining vocabulary in foreign learners' dictionaries. *Lexicography*, 2: 101-119.

HOFLAND, K. and JOHANSSON, S. (1982). *Word Frequencies in British and American English*. Bergen: The Norwegian Computing Centre for the Humanities.

JUILLAND, A. and CHANG-RODRIGUEZ, E. (1964). *Frequency Dictionary of Spanish Words*. The Hague: Mouton & Co.

KILGARIFF, A. (1997). Putting frequencies into dictionaries. *International Journal of Lexicography*, 10(2): 135-155.

KUČERA, H. and FRANCIS, W.N. (1976). *Computational Analysis of Present-Day American English*. Providence: Brown University Press.

LEE, D. (2001). Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, 5(3): 37-72.

THORNDIKE, E. and LORGE, I. (1944). *The Teacher's Word Book of 30,000 Words*. New York: Teachers College, Columbia University.

WEST, M. (1935). *Definition Vocabulary*. Department of Educational Research Bulletin, no. 4. Toronto: University of Toronto.

STUBBS, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Electronic Dictionaries

Cambridge Advanced Learner's Dictionary CD-ROM 3rd Edition. (2008). Cambridge University Press. (*CALD3*)

Collins COBUILD Dictionary on CD-ROM 2006. (2006). HarperCollins Publishers. (*COBUILD5*)

Longman Dictionary of Contemporary English 5th Edition DVD-ROM. (2009). Pearson Education Limited. (*LDOCE5*)

Macmillan English Dictionary CD-ROM 2nd Edition Version 2.1. (2007). Macmillan Publishers Limited. (*MED2*)

Oxford Advanced Learner's Dictionary 7th Edition CD-ROM. (2005). Oxford University Press. (*OALD7*)

Building an OLIF-based lexical database for representing constructions

Marc Luder
University of Zurich

Abstract

This paper demonstrates the implementation of a small monolingual lexical database for German – currently 7,000 entries – for the purpose of manual and automated lexical queries. The lexicon is part of a web based text analysis application that serves to analyze systematically clients' narratives from psychotherapy sessions. The narratives, small stories from everyday life, are conceived as dramaturgically constructed and performed linguistic productions (Boothe 2004). The specific function of the lexicon in this context is to provide the means for the lexical coding of the story vocabulary, which is an important step in the narrative analysis. The lexical database is implemented in the OLIF format; the lexicon entries are conceived as constructions based on a rich set of linguistic and extralinguistic features.

Keywords: multi word expression, construction, pattern, narrative analysis, spoken language, corpora.

1. Introduction

The JAKOB[1] lexicon is a monolingual lexical database for German and contains currently about 7,000 entries, mainly verb entries. When the project started in 2002, the dictionary was designed for single word entries. In 2007, the entry format was adapted for the representation of multi word expressions (MWEs); the data structure was adapted to the Open Lexicon Interchange Format (OLIF)[2], an open source lexicon structure (McCormick 2005; McCormick *et al*. 2004) in XML format. Our implementation maps the XML structure onto a MySQL database and is optimized for manual and automated queries.

The background of the lexicon project is a computerized narrative and text analysis application for coding text with predefined categories called *Narrative Analysis JAKOB*[3] (Boothe 2004). The application allows for a systematic, psycho-dynamically oriented analysis of everyday stories, based on transcripts of psychotherapy sessions. The narratives, small stories from everyday life, are conceived as dramaturgically constructed and performed linguistic productions (Boothe 2004). In this context, the

[1] JAKOB is an acronym for "Aktionen und Objekte" (actions and objects).
[2] http://www.olif.net
[3] Erzählanalyse JAKOB: http://www.jakob.uzh.ch

specific function of the lexicon is to provide the means for the lexical coding of the vocabulary as used in the stories, which is an important step in the narrative analysis. Our interdisciplinary lexicon project is located in the field of psychology, but involves likewise linguistics, corpus linguistics, lexicography, interactional linguistics and conversation analysis. The narrative analysis merges approaches from text and discourse analysis with psychoanalytic concepts. The formal steps of the analysis investigate discourse patterns, lexical choice and story dramaturgy, whereas subsequent interpretation steps seek underlying conflict structures represented by the psychoanalytic concepts of wish, fear and defense. The *Narrative Analysis JAKOB* is a tool for diagnostic and psychotherapy process research.

Stories are compact episodic presentations of personal experience; they display facets of inner conflicts. A short example of an everyday story:

Story "Typisch Frau" – *Typically woman* –

1	ich habe ihm dann später äh noch gesagt. *later I said to him*.
2	äh er habe mich dann schaurig verletzt mit dieser Bemerkung. *that he was hurting me with this remark.*
3	und dann fand er dann aber auch sehr wie mein Vater auch. *and he then found (said) as did my father before*.
4	also was ich jetzt für ein Zeug mache. *thus what I'm making a fuss*.
5	und das sei typisch Frau, solches Zeugs immer auseinanderzubeinen und zu analysieren, oder. *and it is typically woman to always debone and analyze those things.*
6	und schon hatten wir natürlich wieder Krach miteinander, oder. *and yet again we had a tiff together*.

1.1. The computer assisted text analysis application

The procedure of the narrative analysis is as follows. After extracting the story from the transcript, the text is manually segmented; full stops mark the end of a segment. An NLP-tool does POS tagging, morphological analysis and lemmatization for each word in the input string. Afterwards a very shallow parsing/chunking is done; the often fragmented utterances from spoken language do not allow a deeper syntactic parsing. A simple valency pattern with four slots is assigned to each segment, answering the question "who does what how?", or displaying the pattern 'subject, predicator, object, complement/adverbial'. The actants of the story in subject and object position and the actions fullfilled by them are most important for the narrative analysis. Following the segmentation and the allocation of slots, the story vocabulary is automatically coded, based on the codings stored in the lexicon.

2. The JAKOB lexicon

The lexical database in the OLIF format allows for the entry of single words as well as multi word units. We customize slightly the annotation of lexical entries to support a *Construction Grammar* (Croft 2001; Goldberg 1995) approach with fully lexicalized constructions, providing as much as possible syntactic, semantic and pragmatic information (Luder *et al.* 2008). *Constructions* are single or multi word units, including idioms, metaphors and other phraseological and collocational patterns of different types (Granger and Paquot 2008; Villavicencio *et al.* 2005), according to the language use in the transcripts – Spoken language, German with regional dialect characteristics –. The goal of the project is to establish a prototypical lexicon for constructions as pairings of form and meaning with a rich set of linguistic features to disambiguate the query results. Aside from the attributes provided by the OLIF structure, we use data from different dictionaries and lexical resources in order to enrich the lexicographic description of the entries. The lexicon entries include a conceptual category "JAKOB code", selected from our narrative analysis coding system. The main emphasis lies on the actions told and performed by the narrator, *i.e.* on the coding of actants, verbs and verbal constructions.[4] For this reason, we are especially interested in semantic and pragmatic verb classifications of German, as described and proposed by various authors (Čulo *et al.* 2008; Fellbaum 2007; Fellbaum *et al.* 2006; Hanks *et al.* 2006).

One intention of our project is to integrate theoretical perspectives of *conversation analysis* and *interactional linguistics* with perspectives of *construction grammar* and *corpus linguistics*. Speakers use a big inventory of prefabricated patterns for their utterances, ranging from entirely fixed units to looser phraseological constructs (Moon 2008). As put by Hanks, "human beings store in their brains not just words in isolation, but also sets of stereotypical syntagmatic patterns associated with each word" (Hanks 2004: 245). The data background of the lexicon is therefore the language as used in our corpus of therapy conversations.

2.1. Data and methods for lexicon building

The lexical entries are based on spoken language, mostly from psychotherapy sessions. The examples mentioned in this paper come from three corpora representing three different clients, a total of 450 hours of conversation – the corpora cover the whole sessions, not only the narratives. The therapy sessions are video or audio recorded and transcribed, and interesting discourse phenomena – *e.g.* the usage of multi word expressions, with focus on the notion of *construction* – are investigated by methods derived from conversation and discourse analysis (Goodwin and Heritage 1990). Our attention is focused on a functional perspective rather than on linguistic characteristics. After analyzing a construction in several contexts, the findings are verified and further

[4] For example, the coding system for verbs consists of 93 a priori defined verb categories.

refined by means of a corpus query tool (sketch engine/corpus architect[5]). We search for further quotation examples in alternative corpora (*e.g.* internet documents) to approve or refute our hypotheses about the meaning of the construction and to build the new lexicon entry.

2.2. Basic assumptions for lexicon entries

The following assumptions constitute the conceptual background of the JAKOB lexicon structure:

- Lexicon entries are single words AND *multi word expressions*. The focus lies on MWEs.
- A lexicon entry is a *construction*, *i.e.* a pairing of form and meaning, as postulated by the theories of *construction grammar*. The OLIF categories are slightly customized for this purpose.
- Lexicon entries are *prototypes*: hence, variations (*e.g.* word order, optional complements, and extensions) are permitted.
- Lexicon entries are derived from *normal usage* in the underlying corpora. *Normal usage* refers to the relative frequency of a MWE or word pattern, while rare and novel patterns are classified as exploitations of norms, according to the theory of *norms and exploitations* (Hanks 2004; Hanks and Pustejowsky 2005).

3. The lexicon entry

The most important properties of a lexicon entry as used in our project are presented in Table 1. Several OLIF properties, *e.g.* for morphology, translation, administration, are omitted for the sake of clarity.

3.1. Properties for the lexicon entry “ein Zeug machen”[6]

Basic properties include the *canonical form* “machen ein Zeug” – to make a fuss – and the property *head* for multi word expressions, generally the verb for a verbal phrase.[7] There is a convention to put the head of the MWE in front of the entry string, “machen ein Zeug”. MWEs have a *phraseType*: *e.g.* “multi word”, “set phrase”, “collocation”, “idiom”. The *orthographic variant* is an element of the entry (“machen ein Zeugs”), whereas *cross-references* are links to associated entries and characterized by link types like “synonym”, “near-synonym”, “antonym”, “near antonym”, “cause”, *etc*.

The syntactic properties include part of speech (for MWEs the part of speech of the head), the syntactic type (for verbs *e.g.* reflexive, modal, function verb), the transitivity type, verb part and type of auxiliary. The OLIF values of the property

[5] http://ca.sketchengine.co.uk/
[6] JAKOB lexicon online: http://www.jakob.uzh.ch/lexikon/
[7] See McCormick (2005) for more details on OLIF properties.

synFrame (verbs only) are not appropriate for German, instead we use the subcategorization patterns (Satzmuster) from the German dictionary "Der Kleine Wahrig" (Wahrig 2007). Wahrig (2007) offers 76 different sentence patterns, represented by a three-digit number. Examples are:

> pattern 505: (verb) + acc. object + prepositional object ("sie befragt ihn über den Vorfall"; *she is asking him about the event*).
>
> pattern 513: (verb) + acc. object + adverbial/mode ("die Kälte macht den Aufenthalt ungemütlich", *the coldness makes the stay uncomfortable*)
>
> pattern 601: (verb) + subject "es" (it) + dative object ("es reicht mir"; *I've had it*).

OLIF Property[8]	Value	Description
canForm	machen ein Zeug – *to make a fuss*	canonical form
orthVariant	machen ein Zeugs	orthographic variant
crossRef	*e.g.* machen ein Theater – near-synonym	cross-references with types *e.g.* synonym, antonym, *etc.*
ptOfSpeech	verb	part of speech – head of MWE
head	machen	head of MWE
phraseType	idiom	type of MWE
synFrame	500 – verb + AkkO	"Satzmuster" (Wahrig 2007)
synType	function verb	syntactic behavior
semType	act – unspecified activity	semantic type – OLIF
definition	sich kompliziert verhalten, Aufregung produzieren	free text definition
*pattern (CPA)	[[Human]] machen {ein Zeug | ein Zeugs}	corpus based *verb pattern* (Hanks 2008)
subjField	general – therapy discourse	domain, genre – whole discourse
*text type	(*e.g.* narrative, description, argumentation)	internal, linguistic properties – discourse unit
*register	(*e.g.* formal, informal, humorous, ironic)	style, social situation – local patterns
*topic	(*e.g.* medicine, cooking, job, sports)	local discourse topic – local patterns

Table 1. Sample lexicon entry – most important attributes

[8] Properties marked with an asterisk (*) are additional extensions of the original OLIF properties.

An OLIF entry has a value *definition* (a free text input field for semantic description) and different given semantic type categories for verbs, nouns, and adjectives, *etc.* Values for verbs include *achievement*, *unspecified activity*, *emotion*, *event*, *mental-activity*, *perceptive*, *process*, and *sense*. Crucial for the text analysis application is the semantic typing of nouns, especially for the appropriate coding of nouns and the respective pronouns. Examples are *abstract*, *animate*, *human animate*, *aspective*, *concrete*, *information*, *locative*, and *measure.*

As a second approach to semantic and pragmatic understanding, we use an adaption of the *verb patterns*, as introduced by the English *Corpus Pattern Analysis* (CPA; *cf.* Hanks 2008). Thus the verb pattern for "ein Zeug machen" would be:

[[Human]] machen {ein Zeug | ein Zeugs}

The *CPA ontology* (Pustejovsky, Hanks, and Rumshisky 2004) provides a semantic type structure for nouns; the verb patterns allow to integrate the semantic context of a lexicon entry and thereby to specify the meaning by selectional restrictions.

Further semantic information is collected unsystematically from "Der deutsche Wortschatz nach Sachgruppen" (Dornseiff 2004), "Der Kleine Wahrig" (Wahrig 2007), Duden 11 (Scholze-Stubenrecht and Wermke 2008), VALBU (Schumacher *et al.* 2004), and FrameNet (Fillmore *et al.* 2003), if appropriate. The property *JAKOB code* refers to the categories of the JAKOB coding system, as in the following example.

Example: *JAKOB code* for "ein Zeug machen" *(to make a fuss)* = DAR-KAM[9]

The OLIF property *subjectField* represents the knowledge domain to which the lexical entry is assigned, according to the language used in different domains, *e.g. agriculture, audiovisual, aviation, botany and zoology, budgets and accounting*. As the property *subjectField* does not cover the required spectrum of pragmatic description suited for disambiguating polysemous expressions, we extend the OLIF structure with the pragmatic/functional properties *text type*, *register* and *topic*. The property *subjField* marks the larger context (external, nonlinguistic criteria), whereas the *text type* (discourse type, communication type) represents internal, linguistic criteria like "narrative", "description", "argumentation" (Östmann 2005). The values for *register* relate to local patterns of style and social situation, *e.g.* "formal", "informal", "ironic", "humorous" (Lee 2001). Finally, the values for *topic* are very important for local meaning constitution. They denote the subject of the current discourse unit with heterogeneous items like "cooking", "job", "sports", as collected by the analysis of the transcripts.[10]

[9] This MWE is represented by a combination of two codes: DAR is the code for the category *Darstellen (to present)*, KAM represents the category *Kämpfen (to fight)*.

[10] During the analytic process, the *topic* values provide ample indications for meaning disambiguation.

3.2. Building lexicon entries from corpora: Constructions with “Dampf” and “explodieren”

As an example of the impact of corpus frequencies, Table 2 displays some frequencies for constructions with the expressions “Dampf ablassen” (to let off steam) and “explodieren” (to explode). It aims to illustrate the theory of *Norms & Exploitations* (Hanks 2004; Hanks and Pustejowsky 2005). Norms refer to “normal” usage of constructions, *i.e.* the expression is frequent and possibly documented in dictionaries. By contrast, exploitations illustrate a new and creative use of MWE components created in the interaction; the conventional use of the components is still recognizable, but the hearer must infer the meaning. For example, the speaker could replace the noun in the metaphoric expression “Dampf ablassen” by another noun. The lexicographer has to decide whether these expressions have to be lexicalized, whether they are instances of an abstract construction or creative exploitations.

Expression	*client1*	*client2*	*client3*	*norm/expl.*
(fast) verjagen (vor Wut) *(to burst with anger)*	21	0	0	n
explodieren *(to explode)*	38	13	0	n
Kragen platzen *(“that was the last straw”)*	0	2	1	n
Dampf *(steam)*	9	6	4	n
Dampf ablassen *(to let off steam)*	6	4	2	n
Dampf machen *(to put under pressure)*	2	*1	0	n
seinen Dampf loswerden *(to get rid of so. steam)*	*1	0	0	e
Dampf aufladen *(to put under pressure)*	1	0	0	e
Überdruck ablassen *(to discharge pressure)*	1	0	0	e
Frust ablassen *(to discharge frustration)*	1	0	0	e

Table 2. Constructions with "Dampf" and "explodieren"[11]

[11] Data from three corpora: 4.5 million tokens, German and Swiss German.

4. Problems and future work

A first issue is the representation of partly fixed constructions in the canonical form of the lexical entry, as in the example constructions "haben zu [+ verb, infinitive]" (to have to [+ infinitive]) or "das Zeug zu [Anything] haben" (to have what it takes), which could be represented as suggested here with the variable part of the construction in square brackets. For an automated dictionary query in the field *canonical form*, the content of the square brackets must be skipped; possible semantic fillers of the variable slot are defined by the *verb pattern*. This signifies a step in the direction of a dictionary of constructions.

Another important question is whether we can determine the meaning of words and expressions a priori in a dictionary. There are several positions arguing against such an assumption. Interactional linguists advocate the constitution of meaning as emergent; accordingly, speakers create it interactively in the process of discourse (Deppermann and Spranz-Fogasy 2006; Günthner 2007). Meaning depends therefore on communicative and situational conditions. For example, the single word "Zeug" (stuff) is an expression for arbitrary things and concepts which are not exactly defined, the word sense differs in various contexts. An investigation of the use of constructions including the word "Zeug" over the period of a long-term psychotherapy[12] shows evidence of the fact that the specific meaning of these constructions is emerging in the course of time and becomes more and more fixed over time. Evidence suggests that the therapist is apt to adopt the meaning of the construction from the client (Luder 2009). Meaning can nevertheless not only emerge locally and ad hoc; there is a basic or core semantics that can be lexicalized. The basic semantics is vague and denotes the core meaning. It is complemented during interaction and in the context of the discourse situation by local semantics (Deppermann 2006: 167; Imo 2007: 13-14). To a certain extent, this view is also supported by the theory of *norms and exploitations*, as mentioned before (Hanks 2004; Hanks and Pustejowsky 2005).

Hoey (2005) is postulating in his theory of *Lexical Priming*, that words and collocation patterns get their individual meaning by frequent use in specific contexts: "Every word is primed for use in discourse as a result of the cumulative effectives of an individual's encounters with the word" (Hoey 2005: 13). Language use in interaction with others results in shared meaning constitution. The handling of the emergence phenomenon and the question of how to integrate it in a lexicon represents an interesting problem for further studies.

5. Conclusion

There are several tasks and questions to deal with in the further development of the JAKOB lexicon, and just as many in the development of the computerized analysis application.

[12] There are 446 constructions with "Zeug" in this corpus.

Verb Patterns (CPA) must be explored further and the formal concept must be adapted for German. Verb patterns should help to disambiguate meanings of expressions in the sentence context, in interaction with syntactic valency information and semantic properties.

The extended pragmatic categories *subject field*, *text type*, *register*, and *topic* were implemented only recently; future practice will show which of them are useful and whether the concept of four different pragmatic properties is suitable.

We must find an appropriate representation for the *canonical form* of constructions with variable elements, *e.g.* "haben zu [V.INF]" (to have to + INF). Should the property *canForm* contain variable elements in square brackets or should the formal description of the construction be placed in a separate property "construction"?

An advanced challenge for an NLP application is the question of how to *represent the current topic/local context* of discourse, how to store it temporally, and how to adapt it to the changing context conditions of the running discourse. This mechanism would be a step forward to integrate *dynamic meaning constitution* into the analyzing process.

References

BOOTHE, B. (2004). *Der Patient als Erzähler in der Psychotherapie*. Giessen: Psychosozial-Verlag.

CROFT, W. (2001). *Radical construction grammar. Syntactic theory in typological perspective.* Oxford: Oxford University Press.

ČULO, O., ERK, K., PADÓ, S., and SCHULTE IM WALDE, S. (2008). Comparing and combining semantic verb classifications. *Language Resources and Evaluation*, 42: 265-291.

DEPPERMANN, A. (2006). Konstitution von Wortbedeutung im Gespräch: Eine Studie am Beispiel des jugendsprachlichen Bewertungsadjektivs assi. In A. Deppermann and T. Spranz-Fogasy (eds). *Be-deuten. Wie Bedeutung im Gespräch entsteht*. 2nd ed. Tübingen: Stauffenburg-Verlag: 158-184.

DEPPERMANN, A. and SPRANZ-FOGASY, T. (eds) (2006). *Be-deuten: Wie Bedeutung im Gespräch entsteht* (2nd ed.). Tübingen: Stauffenburg-Verlag.

DORNSEIFF, F. (2004). *Der deutsche Wortschatz nach Sachgruppen* (8th ed.). Berlin: de Gruyter.

FELLBAUM, C. (2007). Argument selection and alternations in VP idioms. In C. Fellbaum (ed.). *Idioms and collocations. Corpus-based linguistic and lexicographic studies.* London: Continuum: 188-202.

FELLBAUM, C., GEYKEN, A., HEROLD, A., KOERNER, F. and NEUMANN, G. (2006). Corpus-based Studies of German Idioms and Light Verbs. *International Journal of Lexicography*, 19(4): 349-360.

FILLMORE, C.J., JOHNSON, C.R. and PETRUCK, M.R.L. (2003). Background to Framenet. *International Journal of Lexicography*, 16(3): 235-250.

GOLDBERG, A.E. (1995). *Constructions: A construction grammar approach to argument structure.* Chicago: University of Chicago Press.

GOODWIN, C. and HERITAGE, J. (1990). Conversation Analysis. *Annual Review of Anthropology*, 19, 283-307.

GRANGER, S. and PAQUOT, M. (2008). Disentangling the phraseological web. In S. Granger and F. Meunier (eds). *Phraseology. An interdisciplinary perspective*. Amsterdam: John Benjamins: 27-50.

GÜNTHNER, S. (2007). Zur Emergenz grammatischer Funktionen im Diskurs: wo-Konstruktionen in Alltagsinteraktionen. In H. Hausendorf (ed.). *Gespräch als Prozess. Linguistische Aspekte der Zeitlichkeit verbaler Interaktion*. Tübingen: Gunter Narr: 125-155.

HANKS, P. (2004). The Syntagmatics of Metaphor and Idiom. *International Journal of Lexicography*, 17(3): 245-274.

HANKS, P. (2008). Lexical Patterns: From Hornby to Hunston and beyond. In E. Bernal and J. DeCesaris (eds). *Proceedings of the XIII. Euralex International Congress*. Barcelona: IULA: 89-129.

HANKS, P. and PUSTEJOVSKY, J. (2005). A Pattern Dictionary for Natural Language Processing. *Revue Francaise de Langue Appliquée*, 10(2): 1-19.

HANKS, P., URBSCHAT, A. and GEHWEILER, E. (2006). German Light Verb Constructions in Corpora and Dictionaries. *International Journal of Lexicography*, 19(4): 439-457.

HOEY, M. (2005). *Lexical priming: A new theory of words and language.* London: Routledge.

IMO, W. (2007). *Construction Grammar und Gesprochene-Sprache-Forschung: Konstruktionen mit zehn matrixsatzfähigen Verben im gesprochenen Deutsch.* Tübingen: Niemeyer.

LEE, D.Y.W. (2001). Genres, Registers, Text types, Domains, and Styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, 5 (3): 37-72.

LUDER, M. (2009). Konstruktionen in der Erzählanalyse JAKOB. In G. Grimm, N. Kapfhamer, H. Mathys, S. Michel and B. Boothe (eds). *Erzählen, Träumen, Erinnern. Erträge Klinischer Erzählforschung*. Sonderband Psychoanalyse 2009-2(23). Lengerich: Pabst: 226-237.

LUDER, M., CLEMATIDE, S. and DISTL, B. (2008). Ein elektronisches Lexikon im OLIF-Format für die Erzählanalyse. In E. Bernal and J. DeCesaris (eds). *Proceedings of the XIII. Euralex International Congress*. Barcelona: IULA: 729-735.

MCCORMICK, S.M. (2005). *The Structure and Content of the Body of an OLIF v.2.0/2.1 File*. Available: http://www.olif.net/documents/NewOLIFstruct&content.pdf [19.2.2010].

MCCORMICK, S. M., LIESKE, C. and CULUM, A. (2004*). OLIF v.2: A Flexible Language Data Standard*. Available: http://www.olif.net/documents/OLIF_Term_Journal.pdf [19.2.2010].

MOON, R. (2008). Sinclair, Phraseology, and Lexicography. *International Journal of Lexicography*, 21(3), 243-254.

ÖSTMAN, J.-O. (2005). Construction Discourse: A prolegomenon. In J.-O. Östman and M. Fried (eds). *Construction grammars. Cognitive grounding and theoretical extensions*. Amsterdam: Benjamins: 121-144.

PUSTEJOVSKY, J., HANKS, P. and RUMSHISKY, A. (2004). Automated induction of sense in context. In *COLING 2004, 20th International Conference on Computational Linguistics*, Geneva: 1-7.

SCHOLZE-STUBENRECHT, W. and WERMKE, M. (2008). *Duden 11 – Redewendungen: Wörterbuch der deutschen Idiomatik* (3rd ed.). *Der Duden in zwölf Bänden*: vol. 3. Mannheim: Dudenverlag.

SCHUMACHER, H., KUBCZAK, J., SCHMIDT, R. and DE RUITER, V. (2004). *VALBU – Valenzwörterbuch deutscher Verben.* Tübingen: Narr.

VILLAVICENCIO, A., BOND, F., KORHONEN, A. and MCCARTHY, D. (2005). Introduction to the special issue on multiword expressions: Having a crack at a hard nut: Special issue on Multiword Expression. *Computer Speech & Language*, 19(4): 365-377.

WAHRIG, G. (2007). *Der kleine Wahrig: Wörterbuch der deutschen Sprache.* München: Bertelsmann.

Automatic lexical acquisition from corpora: some limitations and tentative solutions

Cédric Messiant[1], Thierry Poibeau[2]
Laboratoire d'Informatique de Paris-Nord, Laboratoire LaTTiCe

Abstract

This paper deals with lexical acquisition. We take another look at some experiments we have recently carried out on the automatic acquisition of lexical resources from French corpora. We describe the architecture of our system for lexical acquisition. We formulate the hypothesis that some of the limitations of the current system are mainly due to a poor representation of the constraints used. Finally, we show how a better representation of constraints would yield better results.

Keywords: lexical acquisition from corpora, syntactic lexicon, subcategorization frames.

1. Introduction

Natural Language Processing (NLP) aims at developing techniques to process natural language texts using computers. In order to yield accurate results, NLP requires voluminous resources containing various information (*e.g.* subcategorization frames – thereafter SCF, semantic roles, restriction of selection, etc.). Unfortunately, such resources are not available for most languages and they are very costly to develop manually. This is the reason why a lot of recent research has been devoted to the automatic acquisition of resources from corpora.

Automatic lexical acquisition is an engineering task aiming at providing comprehensive – even if not fully accurate – resources for NLP. As natural languages are complex, lexical acquisition needs to take into account a wide range of parameters and constraints (*cf.* mainly the kind of information detailed in the previous paragraph along with frequency information) However, surprisingly, in the acquisition community, relatively few investigations have been conducted on the structure of the linguistic constraints themselves.

In this paper, we want to take another look at some experiments we have recently carried out on the automatic acquisition of lexical resources from French corpora. The task consists, from a surface form, in trying to find an abstract lexical-conceptual structure that justifies the surface construction (taking into account the relevant set of

[1] CNRS and Université Paris 13, cedric.messiant@lipn.univ-paris13.fr
[2] CNRS and Ecole Normale Supérieure, thierry.poibeau@ens.fr

constraints for the given language). Here, in order to get a tractable model, we limit ourselves to the acquisition of subcategorization frames from corpora. The task is challenging since surface form incorporates adverbs, modifiers, interpolated clauses and some flexibility in the order of appearance of the arguments that, of course, should not affect the analysis of the underlying lexical-conceptual structure.

Most approaches, including ours, are based on simple filtering techniques. If a complement appears very rarely associated with a given predicate, the acquisition process will assume that this is an incidental co-occurrence that should be left out. However, as we will see, even if this technique is efficient for high frequency items, it leaves a lot of phenomena aside.

2. Previous Work in Lexical acquisition from corpora

Large corpora and efficient parsers are now widely available for a growing number of languages. So, even though lexical resources are not always available, it is now possible to acquire large lexicons directly from the observation of word usage in corpora, based on the output of surface parsers. Moreover, using automatic acquisition techniques makes it possible to get frequency information associated with lexical entries, which is not possible simply using a manual approach.

Several systems have been built using this approach, for several languages; see, among others, Brent (1993), Manning (1993), Briscoe and Carroll (1997), Korhonen (2002), Schulte im Walde (2002), Messiant (2008) and Messiant *et al*. (2008). The acquisition process is made of three different steps:

1. all the occurrences of the different verbs are grouped together, along with their complements;
2. tentative constructions for each verb are identified, along with their respective productivity (we call these "tentative constructions" since they may contain modifiers, and not only arguments; tentative constructions need to be filtered to give birth to actual subcategorization frames);
3. rare constructions are filtered out, taking as an hypothesis the fact that too few occurrences of a construction is probably the sign of an error in the analysis (or a sign that the construction includes an adjunct).

All the systems are based on these hypotheses, even though they differ as for their parsing model or filtering strategy.

3. A Lexical Acquisition System for French

3.1. Pre-processing: Morpho-syntactic tagging and syntactic analysis

Our system first tags and lemmatizes corpus data using the TreeTagger and then parses it using Syntex (Bourigault *et al*. 2005). Syntex is a shallow parser for French. It uses a

combination of heuristics and statistics to find dependency relations between tokens in a sentence. It is a relatively accurate parser, *e.g.* it obtained the best precision and F-measure for written French text in the recent EASY evaluation campaign.[3]

Below is an example that illustrates the dependency relations detected by Syntex (2) for the input sentence in (1):

(1) *La sécheresse s'abattit sur le Sahel en 1972-1973.*
(*The drought came down on Sahel in 1972-1973.*)

(2)
```
DetFS|le|La|1|DET;2|
NomFS|sécheresse|sécheresse|2|SUJ;4|DET;1
Pro|se|s'|3|REF;4|
VCONJS|abattre|abattit|4|SUJ;2,REF;3,PREP;5,PREP;8
Prep|sur|sur|5|PREP;4|NOMPREP;7
DetMS|le|le|6|DET;7|
NomMS|sahel|Sahel|7|NOMPREP;5|DET;6
Prep|en|en|8|PREP;4|NOMPREP;9
NomXXDate|1972-1973|1972-1973|9|NOMPREP;8|
Typo|.|.|10||
```

Syntex does not make a distinction between arguments and adjuncts; rather, each dependency of a verb is attached to the verb.

3.2. Pattern extractor

The pattern extractor collects the dependencies found by the parser for each occurrence of a target verb. Some cases receive special treatment in this module. For example, if the reflexive pronoun "se" is one of the dependencies of a verb, the system considers this verb like a new one. In (1), the pattern will correspond to "s'abattre" and not to "abattre". If a preposition is the head of one of the dependencies, the module explores the syntactic analysis to find if it is followed by a noun phrase (`+SN`) or an infinitive verb (`+SINF`).

Example (3) shows the output of the pattern extractor for the input in (1).

(3)
```
VCONJS|s'abattre :
Prep+SN|sur|PREP Prep+SN|en|PREP
```

3.3. The Subcategorization Frame builder

The SCF builder extracts SCF candidates for each verb from the output of the pattern extractor and calculates the number of corpus occurrences for each SCF and verb combination. The syntactic constituents used for building the SCFs are the following:

[3] The scores and ranks of Syntex at this evaluation campaign are available at http://w3.univ-tlse2.fr/erss/textes/pagespersos/bourigault/syntex.html#easy

1. `SN` for nominal phrases;
2. `SINF` for infinitive clauses;
3. `SP[prep+SN]` for prepositional phrases where the preposition is followed by a noun phrase (prep is the prepositional head);
4. `SP[prep+SINF]` for prepositional phrases where the preposition is followed by an infinitive verb (prep is the prepositional head);
5. `SA` for adjectival phrases;
6. `COMPL` for subordinate clauses.

When a verb has no dependency, its SCF is considered as `INTRANS`.

Example (4) shows the output of the SCF builder for (1).

(4) `S'ABATTRE+s'abattre ;;; SP[sur+SN] SP[en+SN]`

3.4. The Subcategorization Frame Filter

Each step of the process is fully automatic, so the output of the SCF builder is noisy due to tagging, parsing or other processing errors. It is also noisy because of the difficulty of the argument-adjunct distinction. The latter is difficult even for humans. Many criteria are not usable because they either depend on lexical information which the parser cannot make use of (since our task is to acquire this information) or on semantic information which even the best parsers cannot yet learn reliably. Our approach is based on the assumption that true arguments tend to occur more regularly and more frequently after the verb than adjuncts. Thus many frequent SCFs in the system output are correct.

We therefore filter low frequency entries from the SCF builder output. We currently do this using Maximum Likehood Estimates (Korhonen, Gorrell and McCarthy 2000). This simple method involves calculating the relative frequency of each SCF (for a verb) and comparing it to an empirically determined threshold. The relative frequency of the SCF *i* with the verb *j* is calculated as follows:

$$rel\_freq(scf_i, verb_j) = \frac{|scf_i, verb_j|}{|verb_j|}$$

$|scf_i, verb_j|$ is the number of occurrences of the SCF *i* with the verb *j* and $|verb_j|$ is the total number of occurrences of the verb *j* in the corpus.

If, for example, the frequency of the SCF `SP[sur+SN] SP[en+SN]` is less than the empirically defined threshold, the SCF is rejected by the filter. The Maximum Likehood Estimates filter is not perfect because it is based on rejecting low frequency SCFs, which leads to sometimes reject frames that are indeed correct. Our filter incorporates specific heuristics for cases where this assumption tends to generate too

many errors. With prepositional SCFs involving one prepositional phrase (PP) or more, the filter determines which one is the less frequent PP. It then re-assigns the associated frequency to the same SCF without this PP.

For example, `SP[sur+SN] SP[en+SN]` could be split into two SCFs: `SP[sur+SN]` and `SP[en+SN]`. In our example, `SP[en+SN]` is the less frequent prepositional phrase and the final SCF for the sentence (1) is (5).

(5) `SP[sur+SN]`

Note that `SP[en+SN]` is an adjunct here.

4. Some difficulties with this kind of approach

This approach is very efficient to deal with large corpora. However, some issues remain. As the approach is based on automatic tools (especially parsers) that are far from perfect, the obtained resources always contain errors and have to be manually validated. Moreover, the system needs to get sufficient examples to be able to infer relevant information. Therefore, there is generally a lack of information for a lot of low productivity items (the famous "*sparse problem*").

More fundamentally, some constructions are difficult to acquire and characterise automatically. On the one hand, idioms are not recognised as such by most acquisition systems. On the other hand, some adjuncts appear frequently with certain verbs (*e.g.* verbs like *dormir* 'to sleep' frequently appear with location complements). The system then assumes that these complements are arguments, whereas linguistic theory would say without any doubt that these are adjuncts. Lastly, surface cues are sometimes insufficient to recognize ambiguous constructions (*cf. …manger une glace à la vanille* 'to eat a vanilla ice-cream' vs *manger une glace à la terrasse d'un café* 'to eat an ice-cream outside the café').

5. Some solutions

These issues do not mean that automatic methods are flawed, but that they have a number of drawbacks that should be addressed. The acquisition process, based on an analysis of co-occurrences of the verb with its immediate complements (along with filtering techniques), makes the approach highly functional. It is a good approximation of the problem. However, this model does not take into account external constraints.

5.1. Idioms and light verb constructions

The fact that some phrasal complements (with a specific head noun) frequently co-occur with a given verb is most of the time useful, especially to identify idioms (Fabre and Bourigault 2008), colligations (Firth 1968) and light verb constructions (Butt 2003). On the other hand, the fact that a given prepositional phrase appears with a

large number of verbs may indicate that the preposition introduces an adjunct rather than an argument.

So, instead of simply capturing the co-occurrences of a verb with its complements, we have a number of important features which are available:

- indicator of the dispersion of the prepositional phrases (PPs) depending on the prepositional head (if a PP with a given preposition appears with a wide range of different verbs, it is more likely to be a modifier);
- indicator of the probability for a given PP to appear as an argument rather than as an adjunct (some PPs are rarely arguments, *e.g.* time or location phrases);
- indicator of the co-occurrence of the nominal head of an argument (NP or PP) with a verb (if a verb appears frequently with the same nominal head, it is more likely to form a semi-idiomatic expression);
- indicator of the complexity of the sentence to be processed (if a sentence is complex, its analysis is less reliable). We can calculate a "confidence measure" of the syntactic analysis of a sentence and thus of the syntactic frame extracted from this sentence;
- lastly, semantic typing of the arguments, to distinguish two similar SCFs if they differ only from a semantic point of view.

To be able to do this, the pattern extractor has to be modified in order to keep most of the information that was previously rejected as not relevant. We then need to calculate these indicators so that they can be taken into account.

All these constraints can be evaluated separately, so that we obtain for each of them an ideal evaluation of the parameter. There are two ways of doing this:

1) by automatically inferring the different weights from a set of annotated data

or

2) by estimating the results of various manually defined weights.

We are currently using the second method since data annotation is very costly. However, the first approach would certainly lead to more accurate results. The weight and the ranking of the different constraints must then be examined. A linear model can provide a first approximation but there are surely better ways to integrate the different constraints. Some studies may provide some cues but we still need to evaluate them in our framework (Blache and Prost 2008).

This is the reason why we are interested in constraints models. We assume that language can be represented using a set of constraints, themselves modelled as "dynamic forces". The same idea has been developed in various theories (*e.g.* Shieber 1992; Blache 2001). However, it seems that it has not been fully developed in the case of acquisition processes.

5.2. Manual Validation

The approach requires manual validation. Rather than leaving the validation process apart for further tedious examination by a linguist, we propose to integrate it in the acquisition process itself. Taking into consideration the number of examples and the complexity of the sentences used for training, it is possible to associate confidence scores with the different constructions of a given verb: the linguist is then able to quickly focus on the most problematic cases. It is also possible to propose tentative constructions to the linguist, when not enough occurrences are available for training.

Lastly, when too few examples are available, the linguist can provide relevant information to the machine. However, a well-designed and dynamic validation process makes it possible to decrease by one order of magnitude the time spent on validating the data (Figure 1 presents an overview of the system interface).

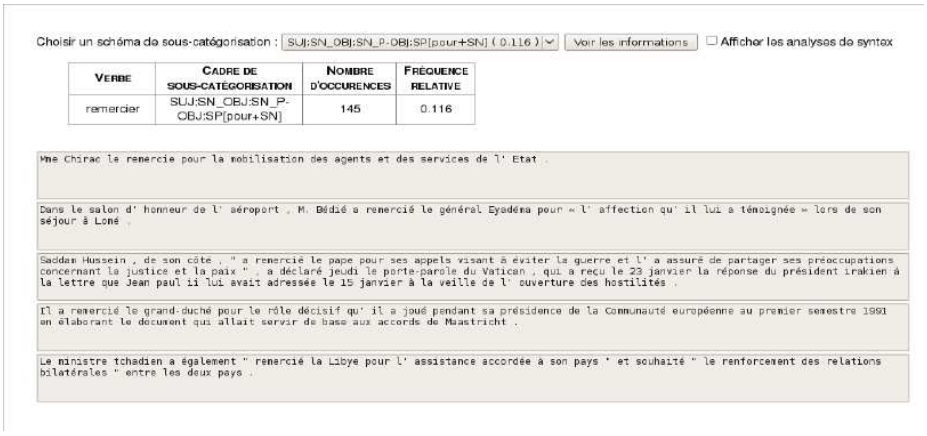


Figure 1. An overview of the interface of the system: http://www-lipn.univ-paris13.fr/~messiant/lexschem.html

6. Conclusion

This paper introduced LexSchem – a large-scale subcategorization frame lexicon for French verbs. The lexicon has been automatically acquired from a large corpus and currently contains 10,928 lexical entries for 5,261 French verbs. The lexicon is provided with a graphical interface and is made freely available to the community via a web page. Future work will include improvement of the filtering module (*e.g.* experimenting with SCF-specific thresholds or smoothing using semantic back-off estimates), automatic acquisition of subcategorization frames for other French word classes (*e.g.* nouns), and automatic classification of verbs using the subcategorization frames as features (Levin 1993).

Acknowledgements

This research was carried out as part of an Alliance grant funded by the British Council and the French Ministry of Foreign Affairs; Cédric Messiant's PhD is funded by a DGA/CNRS Grant.

References

BLACHE, P. (2001). *Les Grammaires de Propriétés: des contraintes pour le traitement automatique des langues naturelles.* Paris: Hermes Sciences.

BLACHE, P. and PROST, J.-P. (2008). A Quantification Model of Grammaticality. In *Proceedings of the 5th International Workshop on Constraints and Language Processing* (CSLP2008). Hamburg.

BRENT, M.R. (1993). From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. *Computational Linguistics*, 19: 203-222.

BOURIGAULT, D., JACQUES, M.-P., FABRE, C., FRÉROT, C. and OZDOWSKA, S. (2005). Syntex, analyseur syntaxique de corpus. In *Actes de la conférence Traitement Automatique des Langues Naturelles* (TALN), Dourdan.

BUTT, M. (2003). The Light Verb Jungle. *Harvard Working Papers in Linguistics*, 9: 1-49.

BRISCOE, T. and CARROLL, J. (1997). Automatic extraction of subcategorization from corpora. *Proceedings of the Meeting of the Association for Computational Linguistics*. Washington: 356-363.

FABRE, C. and BOURIGAULT, D. (2008). Exploiter des corpus annotés syntaxiquement pour observer le continuum entre arguments et circonstants. J*ournal of French Language Studies*. 18(1): 87-102.

FIRTH, J.R. (1968). A synopsis of linguistic theory. In F.R. Palmer (ed.). *Selected Papers of J.R. Firth 1952–59*. London: Longmans, 168-205.

KORHONEN, A. (2002). Subcategorization Acquisition. *Technical Report UCAM-CL-TR-530*. University de Cambridge: Computer Laboratory.

KORHONEN, A., GORRELL, G. and MCCARTHY, D. (2000). Statistical filtering and subcategorization frame acquisition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Hong Kong.

LEVIN, B. (1993). *English Verb Classes and Alternations: a preliminary investigation*. Chicago: University of Chicago Press.

MANNING, C.D. (1993). Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. In *Proceedings of the Meeting of the Association for Computational Linguistics*. Columbus: 235-242.

MESSIANT, C. (2008). ASSCI: A Subcategorization Frames Acquisition System for French Verbs. In *Proceedings of the Conference of the Association for Computational Linguistics* (ACL, Student Research Workshop), Columbus: 55-60.

MESSIANT, C., KORHONEN, A. and POIBEAU, Th. (2008). LexSchem: A Large Subcategorization Lexicon for French Verbs. *Proceedings Language Resources and Evaluation Conference* (LREC). Marrakech.

PRINCE, A. and SMOLENSKY, P. (2004). *Optimality Theory: Constraint Interaction in Generative Grammar*. Oxford: Blackwell.

SCHULTE IM WALDE, S. (2002). A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG. In *Proceedings Language Resources and Evaluation Conference* (LREC). Las Palmas: 1351-1357.

SHIEBER, S. (1992). *Constraint-based Grammar Formalisms*. Cambridge: The MIT Press.

Processing collocations in a terminological database based on a cross-disciplinary study of scientific texts

Mojca Pecman[17], Claudie Juilliard, Natalie Kübler, Alexandra Volanschi
CLILLAC-ARP, Université Paris Diderot – Paris 7

Abstract

This paper presents an ongoing research aiming to build a terminological and phraseological database that would serve for the design of a dictionary-type tool for scientists and translators. This research is carried out at Paris Diderot – Paris 7 University by a group of researchers working on corpus linguistics and LSP within the ESIDIS-ARTES project (Etude des Spécificités et Invariants des DIscours Scientifiques en vue de l'Aide à la Rédaction de TExtes Scientifiques). The project is oriented towards multi-disciplinary discourse analysis of invariants and variations in scientific language with the purpose of designing tools for assistance in scientific writing and translation. The paper discusses the various problems arising from the processing of collocations in a terminological database. We first present the general methodology and principles for encoding collocations in a terminological database. We then discuss the advantages of establishing multiple relations between collocations and show how to encode semantic preference and prosody, and to work towards a more meaning-oriented classification of collocations.

Keywords: phraseology, collocations, terminology, database, semantic preference, semantic prosody, corpus linguistics, assistance in scientific writing, translation, non-native speakers.

1. Introduction

The present communication illustrates the principles and methodology of processing collocations in a terminological database, developed within the ESIDIS-ARTES project. The ESIDIS-ARTES project is carried out at Paris Diderot University by a group of researchers working on corpus linguistics and Language for Specific Purposes (LSP). The project is oriented towards multi-disciplinary discourse analysis of invariants and variations cross scientific domains with the purpose of designing tools for assistance in scientific writing and translation. The main objective of this project is the design of a methodology for investigating scientific cross-disciplinary phraseology through a corpus-based study leading to the creation of reusable linguistic resources. The project involves among others the students working on their Masters

[17] mpecman@eila.univ-paris-diderot.fr

Degree in Language Engineering and Specialised Translation[18] who use a predesigned database for encoding terminological and collocational information through a corpus-based study in a variety of domains (*cf.* Kübler 2003, forthcoming). Our ultimate objective is to make this database accessible online in order to facilitate the editing process and to allow an easy access to encoded information for language users.

The encoding of collocations in a terminological database can provide useful lexical information on the conventionalities of languages for specific purposes. Such resources can be immensely useful to translators, language learners or even to advanced non-native speakers who need to write articles or other text types within the scope of their discipline. Our aim is thus to design a terminological database which provides not only the terminology of a specific scientific field, but also the most frequent collocational patterns in which this terminology appears.

The project aims thus at developing two distinct, but closely linked applications: (1) a comparable French/English corpus containing key scientific documents in different domains (namely earth and planetary sciences, medicine, chemistry, biology, and informatics), and (2) a database intended to receive the linguistic resources, with the possibility of developing a dictionary based tool for the assistance in scientific writing. Both applications serve as bases for enhancing our knowledge on languages functions and features within scientific discourse.

Pursuing this objective, we discuss in this article the various problems arising from the processing of collocations in a terminological database. Several extensive in-depth studies on the lexicographical treatment of collocations may be quoted: Benson *et al*. (1997); Fontenelle (1994; 1997); Hausmann (1979); Heid (1992); Heid and Freilbott (1991); L'Homme (1997; 2007); L'Homme and Meynard (1998); Maniez (2001); Meynard (1997); Pavel (1993); Pecman (2004; 2005; 2008); Siepmann (2006), and Volanschi (2008). Nevertheless the problems that arise from the encoding of collocations and the design of collocational dictionaries are not yet solved and the processing of collocations remains a challenging issue.

Bearing in mind our objective of conceiving a procedure for systematically processing collocations in a database, we will discuss a number of lexicographical problems that arise from such a project. We discuss the principles for encoding collocations in a terminological database and the advantages of establishing multiple relations between collocations. We also discuss how to encode semantic preference and prosody, and the possibilities for more meaning-oriented classification of collocations.

[18] Masters studies at the department of Études Interculturelles de Langues Appliquées (EILA), Paris Diderot – Paris 7 University.

2. General framework

2.1. The role of phraseology in scientific communication

There is an increasing number of scientists who use their second language in their scientific communications, as nowadays scientific literature is almost exclusively written in English. The terminology of a specific field of knowledge is generally well known by domain specialists in both their native language and second language. When students reach the Master's level, they start acquiring the specialised terminology rather quickly. It is highly likely that the necessity to read the literature directly in a second language accelerates the cross-language transfer of specific concepts. On the other hand, the combinatory properties of terms in a second language are often less obvious. All lexical items, including terms, have specific collocational profiles which cannot be literally transferred form the source language to the target language. Nevertheless, this codified phraseology plays an essential role in scientific discourse. We think that designing a terminological database which provides not only the terminology of a specific scientific field, but also the most frequent collocational patterns in which this terminology appears, could be an efficient response to communicative needs of non-native speakers.

In order to ensure this efficiency, the ESIDIS-ARTES database is entirely corpus-based.

2.2. Building corpus-based language resources

The database should provide us with an authoritative description of the terminology and phraseology of scientific language. The data stored in the database should thus be thoroughly corpus-based. With the ARTES project we aim to build a corpus with an online concordancer that would respond to the project expectations of objectivity, accuracy, and reliability. We plan to annotate the corpus texts according to TEI XML format which would allow for more complex searches in the corpus.

At this stage of the project, we use the Web Assisted Language Learning (WALL)[19] concordancer which is an experimental tool allowing us to interrogate data using Perl regular expressions (Kübler and Foucou 1999). Although we have set rigorous criteria for text selection, we are nevertheless confronted to well known difficulties in specialised corpus design: the relative absence of scientific literature in the French language and the overwhelming amount of literature written by non-native speakers. Each text is thus examined thoroughly before being included into the corpus and all necessary measures are taken in order to ensure a comparable English-French corpus.

For the purposes of this study, we have interrogated through WALL several subcorpora representing a number of disciplines of earth and planetary sciences, namely volcanology, seismology, plate tectonics, glaciology and orogeny. These

[19] http://wall.eila.univ-paris-diderot.fr

corpora contain approximately 32 million words (18 million in English and 14 million in French). Nevertheless, with this experimental tool, an even distribution of texts across disciplines could not be guaranteed.

A set of principles have been followed in designing the database and the overall methodology for collecting collocational resources.

3. Principles for encoding collocations

3.1. Simultaneous compilation of phraseological and terminological resources

By compiling simultaneously both resources, terminological and phraseological, we can ensure the completeness of data.

In view of the goals of the ESIDIS-ARTES project, it would be limiting to separate these two resources for two reasons. First, the domain-specific phraseology is generally linked to specific terms. The organisation of phraseological data within the database is thus more systematic if the phraseology is stored in relation to terms. Second, the language users we target with our tool may need information on both types of linguistic units.

We have developed an experimental database using the Access database management system. In this experimental database, each collocation is associated to a term that plays the role of headword in the database.

For instance, the following collocations in the domain of earth and planetary sciences, and more precisely in the domain of glaciology, are all processed under the headword *ice sheet* as a part of its terminological record:

> Term: *ice sheet*
>
> Collocations: *beneath the ice sheet, migration of the ice sheet, retreat of the ice sheet, to be overridden by the ice sheet, ice sheet flows, ice sheet melts, ice sheet slides, ice sheet spreads*

The same collocations can be linked to other terms in the database. For example, the collocation *ice sheet melts* could appear in the terminological record of the verb to *melt*, together with its other potential collocations:

> Term: *to melt*
>
> Collocations: *ice sheet melts, rock melts, crust melts, plume melts*

The ascription of a same collocation to two different terms is handled manually.

3.2. Design of a relational database scheme

Not only is each collocation associated with a term that plays the role of headword in the database, but each collocation is also processed as a distinct secondary entry. This implies, from the point of view of database architecture, that terms and collocations are separated in two distinct tables (as shown in Figure 1). The connection between

terms and collocations is provided by the secondary key field (id_term) which is added to the table of collocations. Each collocation is recorded in a separate entry and can be processed independently of a headword term to which it is linked. In that way, we can link collocations together and regroup them according to their various semantic values: we can, for example, indicate which collocations are synonyms or which collocations are equivalent when working form a bilingual perspective (see part 4). The recording of multiple cross-relations between collocations would be impossible if all the collocations associated to the same headword were listed in the same record (as shown in Figure 2).

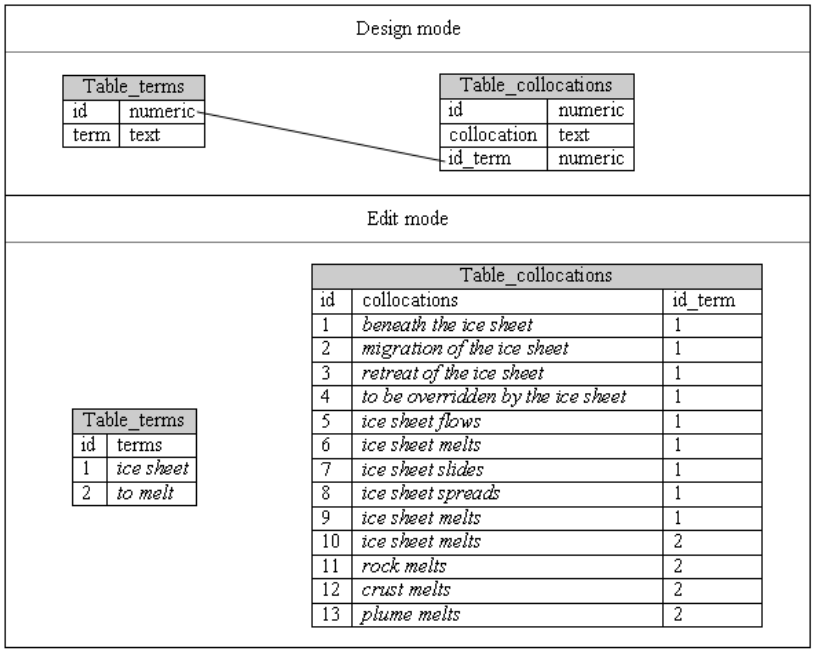


Figure 1. View of the tables Terms and Collocations (adopted scheme)

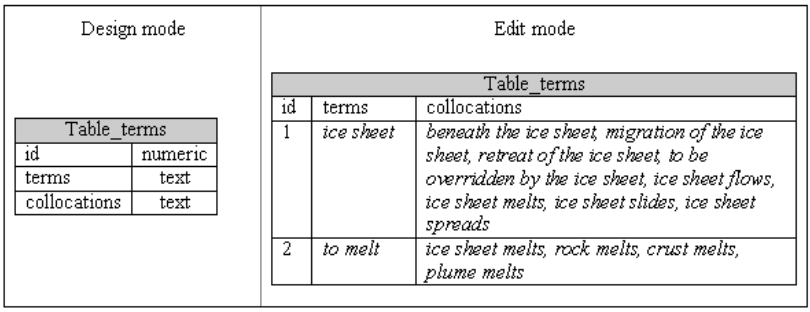


Figure 2. View of the table Terms (non-adopted scheme)

3.3. Normalization of resources

By the term 'normalization of resources' we imply the necessity for systemizing the encoding procedure for all collocations. Collocational resources recorded in the database have to be lemmatised and encoded using appropriate abbreviations.

The following abbreviations have been used so far: ***sth, sb, sb's, to do sth, doing sth*** in English; ***qch, qn, faire qch*** in French. They enable us to extract collocations form discourse and to turn them into abstractions that can be easily inserted into a new context.

> 1. "We largely discount the potential, and speculated, influence of *supersonic transmission velocities, because even though the MSH airwaves might have originated as shock waves (i.e., [23] and [30]), reasonable shocks would decay to sonic speeds within a few kilometers of the source.*"

Example 1 illustrates the importance of normalising resources. The first line: "We largely discount the potential, and speculated, influence of...", offers interesting word combinations: *to discount the potential influence of sth (on sth), to discount the speculated influence of sth (on sth), to largely discount sth.* Once they are lemmatised and correctly encoded, these collocations can be efficiently integrated into new contexts by language users.

Table_collocations

id	collocations	id_term	id_gram
1	*beneath the ice sheet*	1	6
2	*migration of the ice sheet*	1	3
3	*retreat of the ice sheet*	1	3
4	*to be overridden by the ice sheet*	1	7
5	*ice sheet flows*	1	2
6	*ice sheet melts*	1	2
7	*ice sheet slides*	1	2
8	*ice sheet spreads*	1	2
9	*ice sheet melts*	1	2
10	*ice sheet melts*	2	2
11	*rock melts*	2	2
12	*crust melts*	2	2
13	*plume melts*	2	2
14	*retrait de la calotte glaciaire*	3	3
15	*extension de la calotte glaciaire*	3	3
16	*étendue de la calotte glaciaire*	3	
17	*fonte de la de la calotte glaciaire*	3	
18	*amincissement de la calotte glaciaire*	3	
19	*creuser la calotte glaciaire*	3	
20	*calotte glaciaire s'étend*	3	
21	*calotte glaciaire fond*	3	

Table_grammatical_constructions

id	grammatical_construction
1	*vb.* + term
2	term + *vb.*
3	*n.* + *prep.* + term
4	*n* + term
5	*adj.* + term
6	*prep.* + term
7	to be *vb.(p.p.)* by + term

Table_link_type

id	link_type
1	equivalence
2	derivation
3	synonymy
4	antonymy

Table_ collocations_link

Id	id_colloc_1	id_colloc_2	link_type	note
1	3	14	1	
2	6	21	1	
3	8	20	1	
4	16	20	2	
5	17	21	2	

Figure 3. View of the tables designed for storing collocations

4. Establishing the relations between collocations

The ARTES project aims at encoding multiple relations between collocations in order to ensure multiple usages of the data stored in the database. Figure 3 shows the tables which were designed for storing collocations and for recording the information on their grammatical structure and their various semantic relations.

4.1. Intra-linguistic relations

Within a particular language, we intend to encode information on the grammatical structure of collocations, on their derivational capacities, on synonymy and antonymy, as illustrated in Figure 4.

Apart from grammatical structure, the linguistic information that we provide on collocations derive essentially from our capacity to process collocations as separate entries and to link them together according to their various semantic values (see 3.2.).

GRAMMATICAL COMPOSITION
ex. term + vb. : *ice sheet flows, ice sheet melts, ice sheet slides, ice sheet spreads*
ex. n + prep. + term: *migration of the ice sheet, retreat of the ice sheet*

DERIVATIONAL CAPACITIES
ex. *ice sheet melts <> melting of the ice sheet*
ex. *calotte glaciaire fond <> fonte de la calotte glaciaire*

SYNONYMY
ex. *beneath the ice sheet <> under the ice sheet*

ANTONYMY
ex. *beneath the ice sheet <> above the ice sheet, on ice sheet surface*

Figure 4. Linguistic information in relation with collocations provided in the database

4.2. Cross-linguistic relations

The cross-linguistic relations that are recorded in our database are translational equivalences (Figure 5). Like terms, collocations are processed as language units which assure the transfer of meaning from source language to target language.

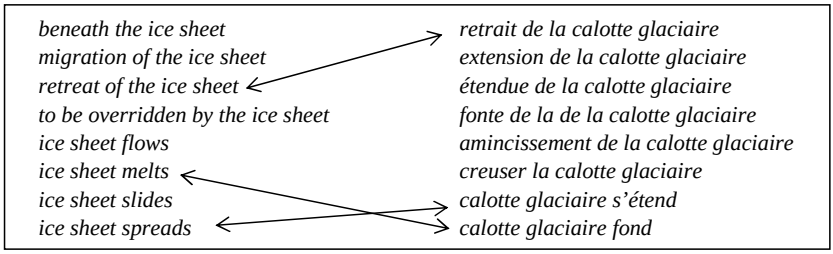


Figure 5. Establishing cross-linguistic equivalences between collocations

In order to avoid too many cross-references when dealing with synonymous collocations, we have recently adopted a more sophisticated scheme which consists in first assigning collocations which bear the same meaning to a set, and then linking equivalent sets across languages. Regrouping semantically related collocations in *synsets* is inspired by the WordNet approach to synonyms, an approach which has been, for instance, successfully integrated in the Dictionnaire d'Apprentissage du Français Langue Etrangère ou Seconde (DAFLES) (Selva *et al*. 2002).

5. How to encode semantic preference and prosody?

Semantic preference is often defined as a collocation between a word and a set of semantically related words. It is thus challenging to attempt to encode semantic preference in our database. Similarly, semantic prosody can be an interesting property of collocations to look at, as it allows us to establish a relation between a lexical item and a set of words which bear negative or positive connotation (*cf*. Kübler *et al.* in print).

The advantages of encoding semantic preferences and prosodies lie in encoding simultaneously a series of lexical items which combine with the same node. Nevertheless, the difficulty of such an approach to linguistic data is that all data has to be categorised beforehand according to semantic relations of hyperonymy and hyponymy or, in the case of semantic prosody, according to their potential connotations, which is a very difficult analysis to conduct systematically on all data. Consequently, we find it preferable to develop a model which can be applied selectively whenever it is necessary to encode semantic preference and prosody.

Examples 2 to 6 establish the semantic preference of the verb ***to accommodate*** which has a tendency to combine with nouns designating some general geological phenomenon, such as ***movement, motion, extension, shortening, deformation***, etc.

> 2. "Volcanoes can cause earthquakes by making bodies of rock **accommodate** the **movement** of magma."
>
> 3. "Earthquake-derived strain rates are low along the Nootka Transform, which **accommodates** relative **motion** between the Juan de Fuca and North American plates."

4. "The zone of maximum convergence rates is well north of the surface trace of the HFF system, which **accommodates** much of the geologic **shortening** of the last few million years. "

5. "Analysis of olivine LPO in naturally deformed peridotites shows that upper mantle **deformation** is generally **accommodated** by slip on both (010)[100] and (001)[100] systems, with dominance of the former. "

6. "At later stages of convergence, as shown in Fig. 11, the centre of the orogen is characterised by maximum exhumation due to erosion and the activity of normal faults, which **accommodate** lateral **extension**."

Examples 7 to 10 corroborate Stubbs's (1996: 173) study on the involvement of strongly negative semantic prosodies in causation. In 40,000 examples of the lemma *cause*, Stubbs reports that its most characteristic collocates are: *accident, concern, damage, death* and *trouble*. In earth and planetary sciences, the verb *to cause* seems to have the same profile as its usual collocates are: *earthquake, landslide, debris flow, collapse, flood, damage*, etc.

7. "Volcanoes can **cause earthquakes** by making bodies of rock accommodate the movement of magma."

8. "However, such a large-release scenario **causes** widespread, unconfined **flooding** (even if the initiation point is moved downslope and the fluid-loss rate is increased up to 1.5 mm/s) that is not consistent with the observed pattern of flow based on where bright sediment deposition actually occurred."

9. "In larger craters, however, gravity **causes** the initially steep crater walls to **collapse** downward and inward, forming a complex structure with a central peak or peak ring and a shallower depth compared to diameter (1:10 to 1:20)."

10. "What are **landslides** and **debris** flows, and what **causes** them? Some landslides move slowly and cause damage gradually, whereas others move so rapidly that they can destroy property and take lives suddenly and unexpectedly."

As Sinclair (1996: 75) already demonstrated in connection with semantic prosody, there is a clear link between semantic prosodies and preferences on the one hand and their role in determining units of meaning on the other. We therefore find it important to design within the ARTES project a model that allows for efficient encoding of semantic preference and prosody in our database. An example of a very simple model for encoding such a type of linguistic information is the use of a specific typographic sign, such as a specific type of brackets (Figure 6).

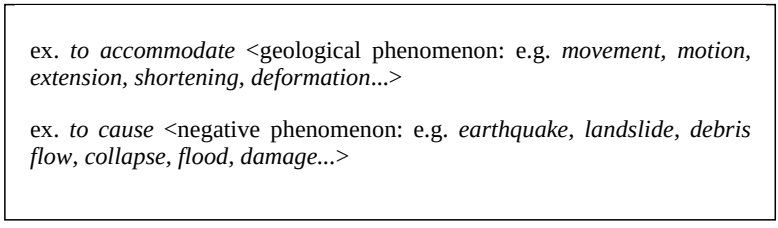


Figure 6. Model for encoding semantic preferences and prosodies

A more sophisticated solution, which we have adopted recently, consists in considering semantic preference and semantic prosody as a possible type of intra-linguistic semantic relation between collocations. The collocations are thus simply assigned a posteriori to one or several sets which group collocations revealing semantic preference (prefset) or semantic prosody (prosset).

6. Meaning-oriented classification of collocations

As partially illustrated with the processing of semantic prosodies and preferences, the objective of the ARTES project is to go beyond the formal description of collocations and to propose a more meaning-oriented classification of collocations. Such approach is particularly interesting when processing collocations belonging to general scientific discourse (Pecman 2007). In scientific articles, the collocations which are not domain-specific are very frequent. They support the discoursal argumentation which is typically coded in scientific communication (cf. Granger and Paquot, this volume).

Examples 11, 12 and 13, also extracted from WALL, show instances of collocations belonging to general scientific discourse.

> 11. "Inundated areas **predicted by** the model **are generally in good agreement with** observations. However, the distal lobe morphology **predicted by** the model **is not in good agreement with** the observed lobe morphology."
>
> 12. "**Field studies of** modern ice sheets and glaciers on Earth have often shown that basal shear stress values lie between 0.5 and 1.5 bars [Nye, 1952a, 1952b], **a surprisingly narrow range** considering the spatial variability of observed basal conditions, including gradients in temperature, meltwater content, basal debris, till rheology, and other variables."
>
> 13. "This application was chosen because it enables us **to make relatively accurate predictions about** the ice margins and the bed topography underneath the ice due to the radial symmetry of impact craters."

Discourse-semantics functions shared by various scientific disciplines can be identified through a detailed semantic analysis of collocational resources (Figure 7).

Domain-free collocations can provide useful information on the lexico-grammatical profile of scientific language. From the point of view of our database architecture, however, there is a necessity for differential processing of terminological collocations belonging to a specific scientific field as opposed to collocations which are characteristic of a more general scientific discourse. The former are necessarily linked to terms, while the latter are associated to one or more predefined discourse-semantic functions, as illustrated in Figure 8.

COMPARISON
ex. *in good agreement with*
ex. *in poor agreement with*

HYPOTHESIS
ex. *predicted by*
ex. *to make predictions about sth*
ex. *accurate predictions*

METHODS
ex. *field studies*

EVALUATION
ex. *surprisingly narrow range*
ex. *accurate predictions*
ex. *relatively accurate*

Figure 7. Model for encoding domain-free collocations according to discourse-semantic functions

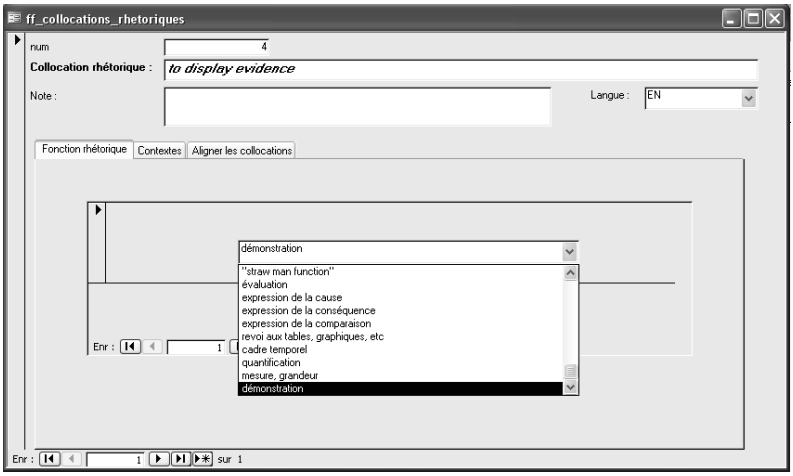


Figure 8. Form for encoding domain-free collocations according to various discourse-semantic functions

7. Conclusion

The database scheme developed within the ESIDIS-ARTES project for encoding collocations in a terminological database points to the importance of encoding multiple information on collocations in order to ensure multiple usages of created resources.

We have shown the possibilities within our database for providing the information on their syntactic structure (ex. vb. + term: *to image microdamage, to detect microdamage*, nom + preposition + term: *accumulation of microdamage, detection of microdamage*, etc.), on their derivational capacities (*e.g. to formulate a hypothesis, formulation of a hypothesis*), on equivalent pairs (*e.g. hypothèse de travail ↔ working hypothesis*), on synonymous pairs (*e.g. to advance a hypothesis, to formulate a hypothesis, to put forward a hypothesis*), or on antonymic pairs (*e.g.. to consolidate a hypothesis* vs. *to invalidate a hypothesis, to refute a hypothesis*). The ESIDIS-ARTES project explores the possibilities for going further in data description. The objective of the project is to go beyond formal description of collocations and to propose a more meaning-oriented classification of collocations, as illustrated with the processing of semantic prosodies and preferences, on the one hand, and of collocations belonging to general scientific discourse on the other. The possibility of providing information on the meaning of collocations is expected to facilitate the presentation and hence the comprehension of the collocational structure of the lexicon.

References

BENSON, M., BENSON, E. and ILSON, R. (1997). *The BBI Dictionary of English Word Combinations*. 2nd ed. Amsterdam and Philadelphia: Benjamins.

FONTENELLE, Th. (1994). Towards the construction of a collocational database for translation students. *Meta*, 39(1): 47-56.

FONTENELLE, Th. (1997). *Turning a Bilingual Dictionary into a Lexical-Semantic Database*. *Lexicographica. Series Maior* 79. Tübingen: Niemeyer.

HAUSMANN, F.J. (1979). Un dictionnaire des collocations est-il possible? *Travaux de linguistique et de littérature*, 17(1): 187-195.

HEID, U. (1992). Décrire les collocations : deux approches lexicographiques et leur application dans un outil informatisé. Colloque international : Phraséologie et terminologie en traduction et en interprétation. *Terminologie et traduction*, 2(3): 523-548.

HEID, U. and FREILBOTT, G. (1991). Collocations dans une base de données terminologiques et lexicales. *Meta*, 36(1): 77-91.

KÜBLER, N. (2003). Corpora and LSP translation. In F. Zanettin, S. Bernardini and D. Stewart (eds). *Corpora in Translator Education*. Manchester: St Jerome Publishing: 25-42.

KÜBLER, N. (forthcoming). Working with corpora for translation teaching in a French-speaking setting. In A. Frankenberg-Garcia (ed.) *New Trends in Teaching and Language Corpora*. Continuum: London.

KÜBLER, N. and FOUCOU, P.Y. (1999). A Web-Based Language Learning Environment: General Architecture. In M. Schulze, M.-J. Hamel, and J. Thompson, (eds) 1999. Language Processing in CALL. ReCALL Special Issue, Hull: 31-39.

KÜBLER, N., PECMAN, M. and BORDET, G. (in print). La linguistique de corpus entretient-elle d'étroites relations avec la traduction pragmatique ?. *Actes des huitièmes journées de Lexicologie, Terminologie, Traduction (LTT) 2009 Lisbonne*. 15-17 October 2009. Lisbonne.

L'HOMME, M.-Cl. (1997). Méthode d'accès informatisé aux combinaisons lexicales en langue technique. *Meta*, 42(1): 15-23.

L'HOMME, M.-Cl. (2007). De la lexicographie formelle pour la terminologie : projets terminographiques de l'Observatoire de linguistique Sens-Texte. in *Actes du colloque BDL-CA* (Bases de données lexicales : construction et applications), 23 avril 2007, OLST, Université de Montréal: 29-40.

L'HOMME, M.-Cl. and MEYNARD, I. (1998). Le point d'accès aux combinaisons lexicales spécialisées : présentation de deux modèles informatiques. *TTR : traduction, terminologie, rédaction*, 11(1): 199-227.

MANIEZ, F. (2001). Extraction d'une phraséologie bilingue en langue de spécialité : corpus parallèles et corpus comparables. *Meta*, 46(3): 552-563.

MEYNARD, I. (1997). Approche hypertextuelle via HTML pour un outil de consignation bilingue des combinaisons lexicales spécialisées. *Actes du Congrès international de terminologie*, San Sebastian, 12-14 novembre 1997, San Sebastian, IVAP/UZEI: 675-689.

PAVEL, S. (1993). La phraséologie en langue de spécialité. Méthodologie de consignation dans les vocabulaires terminologiques. *Terminologies nouvelles*, 10: 23-35.

PECMAN, M. (2004). *Phraséologie contrastive anglais-français : analyse et traitement en vue de l'aide à la rédaction scientifique*. Thèse de doctorat. 9 déc. 2004. Dir. Henri Zinglé. Université de Nice-Sophia Antipolis.

PECMAN, M. (2005). Systemizing the notation and the annotation of collocations. *Jezikoslovlje*. vol. 6, no 1. A linguistic journal published by the Faculty of Philosophy Josip Juraj Strossmayer University Publications. Osijek: 79-93.

PECMAN, M. (2007). Approche onomasiologique de la langue scientifique générale. *Revue française de linguistique appliquée*, 12(2) "Lexique des écrits scientifiques": 79-96.

PECMAN, M. (2008). Compilation, formalisation and presentation of bilingual phraseology: problems and possible solutions. In Granger, S. and F. Meunier (eds). *Phraseology in language learning and teaching*. Amsterdam/Philadelphia: John Benjamins: 203-222.

SELVA, T., VERLINDE, S. and BINON J. (2002). Le Dafles, un nouveau dictionnaire électronique pour apprenants du français. In A. Braasch (eds). *Proceedings of the Tenth EURALEX International Congress on Lexicography*. Copenhagen, 13-17 August 2002. Copenhagen: CST: 199-208.

SIEPMANN, D. (2006). Collocation, Colligation and Encoding Dictionaries. Part II: Lexicographical Aspects. *International Journal of Lexicography*, 19(1): 1-39.

SINCLAIR, J. MCH. (1996). The search for units of meaning. *Textus* IX: 75-106.

STUBBS, M. (1996). *Text and Corpus Analysis*. Oxford: Blackwell.

VOLANSCHI, A. (2008). *Étude et modélisation des phénomènes collocationnels : Implémentation dans un système d'aide à la rédaction en anglais scientifique*, Thèse de doctorat. 5 décembre 2008, Université Paris Diderot. Sous la direction de : Prof. Natalie Kübler.

FDVC
Creating a corpus-driven frequency dictionary of verb phrase constructions for Hungarian

Bálint Sass[1], Júlia Pajzs[1]
Research Institute for Linguistics, Hungarian Academy of Sciences

Abstract

We present a method for creating a special monolingual dictionary. It is a corpus-driven frequency dictionary whose entries are *verb phrase constructions* (VPCs), not words. It does not contain explicit definitions but corpus examples. A significant part of the process is *automatic*: we use not only natural language processing tools to analyse a text but also specific algorithms which take over some of the typical tasks of the lexicographer, that is, we create the raw entries fully automatically. As a result, the manual lexicographic work is reduced, and the dictionary can be created with a low budget. We are creating a Hungarian dictionary but the core methodology is *language independent.* The methodology can be applied for other languages as well. This kind of dictionary can be useful both in language teaching and natural language processing.

Keywords: semi-automatic dictionary writing, automatic entry creation, verb phrase constructions.

1. Introduction

At the beginning of the 21st century one of the burning issues of lexicography is to find ways of automating the process of dictionary creation. This process is notoriously costly and labour intensive but by using today's high performance computers, significant progress has been made in many areas. Linguistic data can be collected from large corpora and can be analysed by using concordancers and different collocation analysis tools. The technical aspects of entry writing are automated by dictionary writing systems (DWSs) which formally and structurally verify the content of the dictionary during the dictionary creating process. But the tasks that require real intelligence – such as identifying senses or writing up definitions – still remain human tasks.

The use of corpora as the main source of linguistic data has become commonplace since the Cobuild dictionary (COBUILD 1987). One important observation in corpus-driven lexicography is that the word is not the best starting point to grasp the meaning, because the meaning is usually realized by a combination of certain words. Multiword

[1] Budapest, Hungary, {sass.balint,pajzs@nytud.hu}

units such as collocations and idioms have a crucial role in language understanding and language generation. They must therefore be represented in the dictionaries in much more detail.

In this study, we present a dictionary creating methodology and its application for a special dictionary. Our methodology is in accordance with the two major trends described above. It uses natural language technology tools extensively, not only for traditional language analysis but also for the *automation* of certain lexicographic subtasks. It handles single-word and *multiword* linguistic units within a single algorithmic framework, and puts multiword units as full-fledged lexemes in the focus of our approach.

The outline of our methodology is the following. First, we extract the raw dictionary directly from a corpus using a specific lexical acquisition method. Then we correct and finalize it by manual lexicographic work. The question is: how far can we get by using automatic tools? How much can we reduce manual lexicographic work? Our work can be considered as a small step towards automated lexicography.

2. Our approach

We follow the Sinclairian approach of corpus-driven (Tognini-Bonelli 2001: 84) lexicography. We take a corpus and "jettison ruthlessly" (Hanks 2008: 219) all verbs and constructions which have zero or low frequency in our corpus. In other words, we take the data from the corpus as is, and we do not allow the lexicographer to add any "missing" constructions. Since we know that "authenticity alone is not enough: evidence of conventionality is also needed" (Hanks 2008: 228), we take the most frequent verb phrase constructions (VPCs) into account and record and display their frequency in the dictionary. We focus on frequent patterns and do not seek to cover all possible meanings and all possible uses (Hanks 2001).

The target is all Hungarian VPCs. These constructions consist of a verb plus zero or more noun phrase (NP) or postpositional phrase (PP) dependents of the verb. Thereby, the VPC category includes all kinds of verb-centered expressions such as verb subcategorization frames, light verb constructions, multiword verbs or completely rigid figures of speech. Considering this full spectrum of expressions we can get an overall picture of the lexicon of the whole language. The dependents, which we call *slot*s, are specified by their surface grammatical relationship to the verb. In English there would be a subject slot, an object slot and different prepositional slots; in Hungarian we have slots specified by different case marks and postpositions. A slot is fixed, if the lemma filling the head of the slot phrase is fixed; a slot is free if its head-lemma can be chosen from a broad word class. VPCs having fixed slots are *multiword verbs*. Free slots usually correspond to valencies.

For example, in the VPC *to get rid of something*, there is a fixed object slot (filled by *rid*) and there is a free *of*-slot, thus this VPC is a valence-bearing multiword verb, just like *to take something into account*. Such complex VPCs are in the focus of our

approach. Since they have both fixed and free slots, they are borderline cases between multiword expressions and verb subcategorization frames. Contrary to common intuition, they are very frequent and therefore cannot be treated marginally. Taking the fixed slot(s) as a part of the (multiword) verb itself, we can treat simple and multiword verbs in the same way: they are both "verbs" and can have valencies. This approach has the significant advantage that different properties (*e.g.* argument structure or frequency) of simple and multiword verbs become directly comparable. Entries in our dictionary are VPCs. The microstructure directly integrates phraseology since the basic units are phrases. In order to form more traditional dictionary entries we arrange the VPCs around a verb in a subsequent step.

On the one hand, the Corpus-driven Frequency Dictionary of Verb Phrase Constructions (FDVC) can be called a "meaningless dictionary" (Janssen 2008: 409) in the sense that it does not contain explicit definitions but merely enumerates the frequent VPCs together with corpus frequencies. However, these dictionaries are efficient as most dictionary users only look up basic information. The dictionary also contains suitable corpus sentences exemplifying the different meanings. Meanings are fairly concrete, as VPCs – being collocations – usually have only one meaning (Yarowsky 1993). In fact, most VPCs are real constructions, "form and meaning pairings" (Goldberg 2006: 3), as they cannot be broken down into smaller units without the loss of meaning. Each VPC represents a pattern of use and can be paired with one sense of its main – simple or multiword – verb. In other words, taking the collocation as the basic lexical unit we get rid of a significant part of polysemy.

3. The dictionary creation process

The automatic phase of the dictionary creation process starts from the morphosyntactically tagged and disambiguated Hungarian National Corpus (Váradi 2002). We need units which contain a verb and its dependent phrases, so the first step is clause boundary detection using a set of rules based on conjunction and punctuation patterns. In the shallow syntactic parsing step we identify the verb and its NP or PP dependents using a cascaded regular grammar. We represent NP and PP slots by their case mark or postposition and their heads. The corpus representation that we have at this point can bee seen in Table 1.

Sentence	*A lány vállat vont.*
English translation	*The girl shrugged her shoulder.*
Representation	verb=von 'shrug' subj=lány 'girl' obj=váll 'shoulder'

Table 1. Representation of an example sentence

The most important step is the lexical acquisition algorithm which collects typical VPCs from a corpus represented as described above. An essential property of the algorithm is that it collects all kinds of typical VPCs, *i.e.* verb subcategorization frames, multiword verbs and those kinds of expressions which are multiwords and have valency at the same time. It is based on cumulative frequency of corpus patterns, with appropriate treatment of fixed and free slots (for details and evaluation, *cf*. Sass 2009b). The main idea is that we initially store all slots and all content words and allow the algorithm to get rid of 1) complete slots when they are not an integral part of a VPC; 2) infrequent content words where they are just occupying a valency slot. The algorithm works as follows:

1. The corpus is represented as illustrated in Table 1. We take all clause representations (verb frames) from the corpus together with their frequency counts. We perform *alternating omission*, *i.e.* we add some variants with free slots to the verb frames in this initial list. From our example sentence – "A lány vállat vont." – we obtain the following four verb frames:

   - `verb=`von `subj=`lány `obj=`váll length: 4 (original)
   - `verb=`von `subj obj=`váll length: 3 (subject omitted)
   - `verb=`von `subj=`lány `obj` length: 3 (object omitted)
   - `verb=`von `subj obj` length: 2 (both omitted)

   This step makes it possible to have VPCs with free slots in the final outcome.

2. We sort the obtained verb frame list according to *length* – which is the number of slots plus the number of fixed content words in slots.

3. Starting with the longest one we go through the verb frame list. We discard verb frames with a frequency lower than 5, and add their frequency to a *one-unit-shorter* frame on the list which *fits* it. According to our definition of length, frame S is one-unit-shorter than frame L if S has a free slot where L has a fixed slot, or S has one less free slot. S fits L means that the slot-set of S is the subset of the slot-set of L, and where L has a fixed slot S does not have a different content word.

4. VPCs are the final remaining verb frames, ranked by cumulative frequency.

In our example, the word "váll" frequently occurs in the object slot and the words appearing in the subject slot are more variable. The desired construction – `verb=`von 'shrug' `subj obj=`váll 'shoulder' – fits several sentences with a variety of different subjects, so it can inherit and sum up their frequencies (see step 3 above). Consequently, it will be around the top of the final outcome list with high frequency.

It should be emphasized that in contrast with the earlier processing steps, this algorithm substitutes a certain time-consuming sub-task of the lexicographer which is usually carried out manually by looking up concordances. The algorithm does not

simply query the corpus but organizes and summarizes information extracted from the corpus.

After we have extracted all typical VPCs we arrange them around verbs, and collect example candidates from the corpus automatically for each VPC by taking the ten most frequent clauses which the VPC fits.

4. Manual lexicographic work

As a result of the automatic phase described in the previous section, raw entries, *i.e.* extracted VPCs arranged around verbs with frequencies and examples, are presented to the lexicographer in an XML format. Our aim was to cover VPCs of about 3000 verbs. Setting the frequency threshold on the VPC list to 250, we obtained 8519 VPCs of 2969 verbs at the end of the automatic phase. This constitutes the raw dictionary. The type distribution of these VPCs is shown in Table 2.

type	Example		count	%
1 free slot	*hisz –bAn*	*believe in*	3181	37%
2 free slot	*ad –nAk -t*	*give* OBJ_I OBJ_D	1563	18%
bare verb	*történik*	*happen*	1469	17%
1 fixed slot	*von váll-t*	*shrug shoulder*	1080	13%
1 fixed + 1 free	*vesz -t figyelem-bA*	*take OBJ into account*	1043	12%
other			183	2%
total			8519	100%

Table 2. The type distribution of VPCs in the raw dictionary

There are obviously cases when erroneous or non-existent VPCs are present because of some errors in the automatic processing. First of all, the lexicographer needs to correct these kinds of error. He or she can simply delete the erroneous VPC (or move it to the correct verb, if only the verb-stem was mistaken). A proper evaluation of the automatic phase determines the ratio of how many suggested VPCs were accepted by the lexicographer. Evaluating the first part of our work we found that 2478 out of 2712 (91.4%) were accepted, which is a fairly good result. We can say that the automatic raw entry creation is of good quality as the lexicographers only rarely came across erroneous VPCs.

The second task is to select a suitable example sentence for every VPC which illustrates its meaning well. When doing this, the suggestions made in Kilgarriff *et al.* (2008) are taken into account – choosing full-sentence examples, or at least clauses with full predicate structure, avoiding personal names, etc. Sometimes none of the example sentences collected automatically are correct or appropriate for illustration. In

such cases, other sentences are retrieved from the Hungarian National Corpus by a special corpus query system (Sass 2008).

Finally, the lexicographer decides whether a VPC is an idiom or not. If a VPC with a fixed slot has its own separate meaning, then it is an idiom; but when the word in the fixed slot is just a frequent word occupying a valency position, then it is not an idiom. Idiomatic and non-idiomatic VPCs will be shown in a different way in the final dictionary.

An important aspect of our approach is that raw entries are created fully automatically. This is decidedly different from today's standard process of dictionary creation, where the lexicographer (1) uses a corpus query and/or collocation analysis tool, then (2) lays out the entry in a DWS, (3) copy-pastes the data needed for the entry, and (4) edits the entry in the DWS. Taking the corpus-driven principle seriously we leave some lexicographical decisions to our lexical acquisition tool. The tool collects all information needed for the entry from the corpus automatically, so in the above model it performs the first three steps, the only thing which remains for the lexicographer to do is to edit the entry in an XML editor and produce its final form.

Although our raw entries are of good quality, they are far from perfect. This can be regarded as a drawback of our approach, but we also have the advantage of being able to speed up the dictionary creation process significantly. Using our methodology, the task of the lexicographer is considerably easier and this methodology makes it possible to create smaller budget dictionaries. The programming and support costs – the automatic phase – are estimated to 1 man-year, and the lexicographic work – the manual phase – is also about 1 man-year for a dictionary containing about 3000 verbs and 8000 VPCs altogether. The quality control of dictionaries is indispensable and relatively time-consuming: within the lexicographic work the first pass is about 6 man-months and the control pass is also about 6 man-months.

5. Final form of the dictionary

Beside the traditional (alphabetically verb-ordered) presentation we plan to have several indexes. All of them can be generated automatically:

1. aggregated list of all VPCs sorted by frequency;
2. an index of dependent combinations;
3. an index of the words in fixed slots;
4. an index of number of fixed/free slots;
5. a frequency list of verb stems.

Here is an example entry for the verb *elver* (*to beat*) in XML form. It is in its post-automatic stage, amended by manually choosing one corpus example from the automatically collected example set for each VPC.

```
<entry>
<verb lemma="elver" freq="744"/>
<pattern freq="284">
<frame><p c="-t"/></frame>
<type str="1:01" len="1" fixed="0" free="1"/>
<cits>
  <cit>hogy minap elvertelek azért,</cit>
</cits>
</pattern>
<pattern freq="95">
<frame><p c="-n"/><p c="-t" l="por"/></frame>
<type str="3:11" len="3" fixed="1" free="1"/>
<cits>
  <cit type="sentence">egy pár túlbuzgó helyi tanácselnökön
                       verjék el a port.</cit>
</cits>
</pattern>
<pattern freq="36">
  <frame><p c="" l="jég"/><p c="-t"/></frame>
  <type str="3:11" len="3" fixed="1" free="1"/>
  <cits>
    <cit type="sentence">Már ahol a jég nem verte el a
                         termést!</cit>
  </cits>
  </pattern>
</entry>
```

The corresponding dictionary entry which shows the three most important verb phrase constructions of this verb is the following (frequency values are given between angle brackets):

elver [744]

elver -t [284] hogy minap elvertelek azért, ...

elver -n **por**-t [95] IDIOM egy pár túlbuzgó helyi tanácselnökön verjék el a port.

elver **jég** -t [36] IDIOM Már ahol a jég nem verte el a termést!

The English version of the entry is only for illustration purposes:

beat [744]

beat OBJECT [284] that I beat you yesterday, because ...

beat ON **dust**-OBJECT [95] IDIOM to blame some overzealous local mayors.

beat **ice** OBJECT [36] IDIOM Just where the crop has not been destroyed by the hail!

VPCs (on the left) are translated word-for-word while example sentences (on the right) have free/idiomatic translations, so comparing the two it can be seen that when something is being *destroyed by hail* Hungarians say *the ice beats* something; and *to blame somebody* is expressed in Hungarian by something like *to beat the dust on somebody*.

6. Language independence

It should be noted that our corpus representation seems to be language independent. In fact, it only depends on whether clauses in the given language consist of a verb plus its dependents and whether the relationships between the verb and the dependents can be grasped somehow – *e.g.* by case marks and postpositions in Hungarian (or by word order and prepositions in English). As can be seen in Table 1, our representation only reflects these relationships. The only condition for operating our automatic VPC acquisition algorithm is that it requires a corpus in our representation format.

Accordingly, we expect that our methodology could be applied to other languages. This idea is confirmed by an experiment performed in the Danish language. Using the 300,000 word Danish Dependency Treebank we showed that the representation can be worked out straightforwardly (Sass 2009a). Although this corpus is small for our purposes, running our lexical acquisition algorithm on it provided some promising results, as the following two raw entries show.

se	*komme*
se [28] ‘look’	*komme* [21] ‘come’
se på [9] ‘look at’	*komme* til [11] ‘come in’
	komme i [11] ‘come in’
	komme på [9] ‘come on’
	komme til at [8] ‘be going to’

To sum up, we can say that the dictionary creating methodology described in this paper can be applied to many languages, if we have a suitably analysed corpus or NLP tools (POS tagger, shallow syntactic parser, etc.) to create one.

7. Conclusion

We described the creation of a Corpus-driven Frequency Dictionary of Verb Phrase Constructions (FDVC) for the Hungarian language. We collected all VPCs from a corpus and created raw entries fully automatically and presented them to the lexicographer in a convenient XML form. Thereby the manual lexicographic work has significantly been reduced. Core algorithms are language independent. Using this methodology we can obtain a dictionary which is useful for linguists and grammar writers. It is also a good learner’s dictionary for advanced learners, a dictionary which lists all typical VPCs and “helps students to write and speak idiomatically” (Hanks 2008: 219). Beyond that, it is a rich lexical resource from which many natural language processing tasks could benefit, from information retrieval to machine translation as well.

Acknowledgment

This work was partly supported by the NK-78074 project of OTKA – Hungarian Scientific Research Fund.

References

COBUILD (1987). *Collins Cobuild English Language Dictionary.* London: HarperCollins publishers.

GOLDBERG, A.E. **(**2006). *Constructions at Work.* Oxford: Oxford University Press.

HANKS, P. (2001). The Probable and the Possible: Lexicography in the Age of the Internet. In S. Lee (ed.). *Proceedings of AsiaLex 2001.* Seoul, Korea: Yonsei University: 1-15.

HANKS, P. (2008). The Lexicographical Legacy of John Sinclair. *International Journal of Lexicography*, 21(3): 219-229.

KILGARRIFF, A., HUSÁK, M., MCADAM, K., RUNDELL, M. and RYCHLY, P. (2008). GDEX: Automatically Finding Good Dictionary Examples. In E. Bernal and J. DeCesaris (eds). *Proceedings of the XIIIth EURALEX International Congress.* Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra: 425-432.

JANSSEN, M. (2008). Meaningless Dictionaries. In E. Bernal, J. DeCesaris (eds). *Proceedings of the XIIIth EURALEX International Congress.* Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra: 409-420.

SASS, B. (2008). The Verb Argument Browser. In P. Sojka, A. Horák, I. Kopecek and K. Pala (eds). *Proceedings of the 11th International Conference on Text, Speech and Dialogue.* Heidelberg/New York: Springer: 187-192 (LNCS, vol. 5246).

SASS, B. (2009a). Verb Argument Browser for Danish. In K. Jokinen and E. Bick (eds). *Proceedings of NoDaLiDa 2009.* NEALT: 263-266.

SASS, B. (2009b). A Unified Method for Extracting Simple and Multiword Verbs with Valence Information and Application for Hungarian. In N. Nicolov, K. Boncheva and G. Angelova (eds). *Proceedings of RANLP 2009.* Borovets: 399-403.

TOGNINI-BONELLI, E. (2001). *Corpus Linguistics at Work.* Amsterdam/Philadelphia: John Benjamins.

VÁRADI, T. (2002). The Hungarian National Corpus. In M.G. Rodriguez, C.P.S. Araujo (eds). *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002).* Las Palmas: 385-389.

YAROWSKY, D. (1993). One Sense per Collocation. In *Proceedings of the workshop on Human Language Technology.* New Jersey: Princeton: 266-271.

The Dici project: towards a dictionary of Italian collocations integrated with an online language learning platform

Stefania Spina[1]
University for Foreigners Perugia

Abstract

This paper describes the *Dictionary of Italian Collocations* (*Dici*), a tool based on natural language processing technologies that aims to support foreign language learning activities. More specifically, the *Dici* is designed to be integrated with an online learning environment: in a specific area of an online platform, devoted to the study of vocabulary, students of Italian as a second language can perform receptive and productive learning activities concerning the recognition and the active use of collocations, with the support of all the information stored in the *Dici*. The paper describes the process of extraction of collocations from a reference Italian corpus, the creation of the dictionary, its structure and its integration with the online learning evironment.

Keywords: collocations, dictionary, online learning environment.

1. Introduction

The importance of multi-word units is widely recognized in several domains, such as natural language processing (Smadja 1993; Evert 2005; Tschichold 2008), lexicology (Ježek 2005), lexicography (Benson 1990; Cowie 1981), language acquisition (Nesselhauf 2005) and foreign language teaching (Lewis 2000; Nation 2001; Meunier and Granger 2008). Choosing from among the numerous definitions of multi-word units, here I propose to define them as "complex lexemes that have idiosyncratic interpretations that cross word boundaries" (Sag *et al*. 2002: 2).

Despite the lack of consensus on the notion of collocation and on the criteria for their identification, collocational competence is generally recognized as playing a key role in the linguistic competence of native speakers. As far as second language learners are concerned, it has been stressed (Nesselhauf 2005: 2; Sung 2003) that collocational competence plays a major role in enhancing their fluency – since it provides ready-to-use "chunks" of language – and in making their comprehension easier.

While native speakers have the ability to recognize such combinations as the most appropriate way of expressing a given concept, learners who lack collocational

[1] Department of Language Sciences, University for Foreigners Perugia, Italy, sspina@unistrapg.it

competence often produce unacceptable combinations; this notably happens with collocations,[2] that are based on conventional linguistic habits rather than on semantic restrictions (*fare una domanda*, 'to make a question'). Previous research on collocations in learner language shows that a third of the collocations used by learners are deviant, and that a longer exposure of the learners in native-speaking countries is more effective to gain collocational competence than activities held in language classrooms (Nesselhauf 2005: 237).

This paper is divided into three different sections; in section 2 I briefly present the project of the constitution of *Dici*, a corpus-based *Dictionary of Italian Collocations*. Section 3 deals with the description of the dictionary structure, and will mainly focus on the methodologies of extraction of collocations from corpora and on the different types of information added to the dictionary entries. Section 4 describes the integration of the dictionary with an online learning environment.

2. The *Dici* project

The *Dictionary of Italian Collocations* is part of *Lele* (*Linguistically-enhanced Learning Environment*), a set of online tools that aim to support second language acquisition. The aim of the on-going project is to enrich an online learning environment with linguistic tools created from natural language processing technologies.[3]

The growing availability of natural language processing tools represents a great opportunity for online learning environments, which can benefit from large amounts of structured linguistic data, typically extracted from corpora and stored in databases or dictionaries, and from innovative computational methodologies (Aldabe *et al.* 2006; Tschichold 2006).

The dictionary is therefore conceived as a lexical database that aims to support the processes of training, testing and improving the collocational competence of students of Italian as a foreign language. It is a corpus-based dictionary: multi-word units are extracted from a balanced corpus of Italian language. It relies on statistically-oriented methodologies – multi-word units are automatically extracted and sorted by mixing distinct criteria such as frequency, dispersion in the different textual genres represented in the corpus – as well as on a "phraseological approach" (Nesselhauf 2005: 12-18; Granger and Paquot 2008: 28-29), by which collocations are not only combinations of words occurring with a given frequency, but also lexical items that can be differentiated on the basis of their syntactic and semantic features.

[2] The term "collocation" is used in this article as a hyperonym for all the different typologies of multi-word units, as well as a hyponym, with the specific meaning of "restricted word combination".

[3] The online learning environment, specifically devoted to Italian as a foreign language students, is named April (Ambiente Personalizzato di Rete per l'Insegnamento Linguistico). It is described on the web site http://april.unistrapg.it/.

The dictionary includes lexical combinations with different degrees of cohesion, ranging from completely frozen associations, which undergo no variation (*ferro da stiro* "iron"), to lexical collocations composed of words that tend to co-occur arbitrarily, although with a low degree of fixedness (*fare la doccia* "to take a shower", that can be interrupted by other lexical items, as in *fare una bella/lunga doccia, fare una doccia calda* "to take a long/hot shower", or undergo syntactic transformations such as left dislocations, as in *la doccia l'ho fatta io* "the shower, I took it").

The *Dici* has two main objectives:

- to support research on multi-word units and their computational and lexicographic treatment;
- to provide a natural language processing resource capable of supporting foreign language teaching activities.

Being a learner-oriented tool, the *Dici* is based on a list of the most common Italian collocations, classified according to frequency. A part of the project is also devoted to the attribution of each of the collocations included in the dictionary to a specific level of competence of Italian as a foreign language: frequency alone would not be sufficient to accomplish this goal.

An accurate analysis of evidence coming from learner corpora on the real use of collocations by L2 learners can shed light on the possible connection between collocations and specific levels of competence; in other words, second language acquisition research on how non-native speakers learn word combinations in a foreign language can give an insight into a language teaching issue such as the distribution of word combinations at different levels of competence in syllabus and teaching materials design (Granger 2004: 134-137). This will be one of the future goals of the project.

3. The *Dictionary of Italian Collocations*

3.1. Extraction methodology

The *Dici* is composed of a list of lemmata, which in this case are multi-words units.

The extraction of candidate word combinations has been carried out in a multi-genre, POS-tagged and lemmatized corpus of Italian: the *Perugia Corpus* (PEC). The corpus size is 18 million words, divided into 7 sections representing different text typologies (fiction, non-fiction, press, blogs, academic prose, language of administration and spoken text).[4]

The automatic extraction of candidate word combinations is preceded by the analysis of existing co-occurrence lists. Lists of multi-word units extracted from a spoken Italian corpus (De Mauro *et al.* 1993) and from the *Italian Wordnet* (Roventini *et al.*

[4] The *Perugia Corpus* (PEC), a reference corpus of Italian, is described on the web site http://elearning.unistrapg.it/corpora/.

2000) have been analyzed and manually tagged by their grammatical structure. This manual tagging was aimed to identify all the possible POS sequences; we have only considered the combinations that "convey a content message" or "referential phrasemes" according to Granger and Paquot's typology (Granger and Paquot 2008: 42).

The total of 150 sequences has been sorted by frequency, and the resulting 10 most frequent POS sequences – which cover 75% of all the referential phrasemes of our list – are automatically extracted from the *Perugia corpus*. The 10 selected sequences are listed in Table 1.

ADJ ADV NOUN	*nudo come un verme* "as naked as a worm"
ADJ CONG ADJ	*bianco e nero* "black and white"
ADJ NOUN	*terzo mondo* "third world"
NOUN ADJ	*cassa comune* "common fund"
NOUN CONG NOUN	*andata e ritorno* "back and forth"
NOUN NOUN	*caso limite* "borderline case"
NOUN PRE NOUN	*abito da sera* "evening dress"
VER ADJ	*stare zitto* "keep quiet"
VER ART NOUN	*fare la doccia* "take a shower"
VER NOUN	*voltare pagina* "turn over a new leaf"

Table 1. The 10 POS sequences extracted from the Perugia corpus

This POS-based selection was a first step in filtering all the possible word combinations. In fact, it has been demonstrated that POS tagging and lemmatisation produce a higher level of effectiveness in the automatic extraction of phraseological units (Pazos Bretaña and Pamies Bertrán 2008: 400). After this phase, all the occurrences of the 10 POS sequences have been automatically extracted from the corpus.

The extraction of candidate word combinations is followed by a selection made in three steps:

1. a first selection is based on frequency (all candidates with frequency equal to 1 have been excluded).
2. a second selection is based on dispersion in the 7 corpus sections. The value of pure frequency is re-calculated on the basis of the distribution of a multi-word unit within the 7 sections of the corpus, obtaining a reduced frequency. As a statistical measure of dispersion, the Juilland's D coefficient has been chosen (Juilland and Traversa 1973; Bortolini *et al*. 1971: 23-24).
3. a third selection is based on the manual exclusion of non-collocations: fully compositional and predictable combinations, such as *è un ragazzo* "he is a boy", or part of a larger word combination.

The result of these different levels of selection – POS-based, frequency-based, dispersion-based and manual filtering – is the final list of the dictionary entries.

3.2. Structure of the dictionary

This final list of multi-word units is then enriched with other structured information connected with lexical, syntactical, semantic, contextual and statistical aspects of collocations. This information is stored in a lexical database, which contains the following fields:

- Grammatical category of the multi-word units.
- Internal syntactic description of multi-word units. Data must be grammatically analyzed to allow the recognition of collocations in a text. Moreover, learner errors could be connected to the syntactic configuration of the collocation rather than to its lexical features; for example using the definite article in *fare la confusione*, instead of the correct form *fare confusione* "create confusion"; or omitting the indefinite article in *fare giro*, instead of using the correct form *fare un giro* "take a tour".[5]
- Definition, extracted from existing dictionaries.
- Context (samples of authentic text in which the word combination occurs, extracted from corpora).
- Syntactic variation of multi-word units: it must be specified if they are completely invariable or if they can be interrupted by elements not belonging to the word combination such as adverbs or adjectives (Heid 2008: 345-346). This information is crucial for the retrieval of collocations in new texts.
- Frequency in the whole corpus.
- Frequency in different corpus sections.
- Reduced frequency (dispersion in different corpus sections).

4. Integration with learning environment

It is generally recognized that the role of multi-word units has often been neglected in language pedagogy (Kennedy 2008: 36-40); this may be considered a paradox, as the first studies on word associations just came to light thanks to the work of English teachers in the 30s (Sinclair *et al*. 2004: ix).

In this sense, corpus-based research can offer a valuable contribution to language pedagogy in the field of collocations, in both the aspects of providing data on

[5] The two examples are taken from the *Spoken Corpus of Italian as a second language*, of the University for Foreigners Perugia (Atzori *et al*. 2009; http://elearning.unistrapg.it/osservatorio/Corpora.html), and specifically from the German learners section.

frequency, distribution and use of word associations and of giving teachers deeper insights into the process of language learning (Kennedy 2008: 37).

A great amount of research has been conducted in the last few decades in the field of multi-word units, specially related to dictionaries of collocations (Santos Pereira and Mendes 2002; Alonso Ramos 2003, just to mention two recent examples for languages other than English). In this context, the specificity of the *Dici* project is its integration within an online learning environment. This means that in a specific area of our learning platform, devoted to the study of vocabulary, students of Italian as a foreign language can perform receptive and productive learning activities concerning the recognition and the active use of collocations, with the support of all the information stored in the *Dici*. Some of the features of the dictionary in its integration with the online platform are:

- to automatically recognize and highlight multi-word units in written Italian texts (receptive level);
- to provide a second language guided writing tool, in order to train the collocational competence (productive level);
- to generate collocation tests aimed at assessing the collocational competence of second language students (Jaén 2007).

4.1. How it works

The lexical database is the core of the online linguistic tool and is connected to a POS tagger.[6] The POS tagger first analyzes the input text as a sequence of tokens associated with a POS and a lemma, and then combines the appropriate sequence that forms a collocation, on the basis of the syntactic information provided by the database.

The starting point is a written text inserted within a web page of the learning environment; each student can filter texts by type of vocabulary. In this case, the choice will be by collocations, but other lexical categories – like academic vocabulary (Spina 2009) – can be processed by the *Lele* system.

When a text is filtered, it is first of all tagged and lemmatized, producing a three column list for each word of the text, composed of word - POS - lemma. This phase is invisible to the end user, as well as the second phase, in which the lemma-POS sequences are compared to all the lemma-POS sequences of the database. If an exact match is found, the corresponding word sequence is highlighted in the filtered version of the text, which is finally visible to the student (Figure 1). Clicking on the highlighted word sequences opens a new, small window that reveals further information connected with the selected collocation.

[6] The POS tagger used for the Dici project is TreeTagger (Schmid 1994); the associated Italian parameter file has been developed at the University for Foreigners Perugia. (http://elearning.unistrapg.it/TreeTagger.html).

Figure 1. The integration of Dici with an online learning environment

The combined use of the POS tagger and the database allows for the correct recognition of sequences of words that are stored in the database as collocations. The system can recognize appropriate word combinations even though they are interrupted by items that do not belong to the combination, *e.g.* adverbs or adjectives. Figure 1 illustrates this with the example *facevano una lunga pausa* "they made a long pause".

This database – POS tagger integration is also useful for other tools provided by our linguistic environment, such as an automatic cloze-generator, which produces a cloze test for any given text, hiding all the collocations or just those belonging to a selected grammatical category.

5. Conclusion

The *Dici* is a lexicographic database with two main specific features:

- it is designed to be used for foreign language learning; frequency is thus one of the main criteria for the selection of word combinations;
- it is designed to be integrated with an online learning environment, specifically devoted to language activities (comprehension, production and testing).

Due to this specificity, the *Dici* is far from providing a complete list of Italian multi-word units (even presuming that this would be possible). The complex nature of this project is due to several reasons, including the vagueness of the linguistic status of

word combinations, at the intersection of lexicon and grammar, and the lack of well defined criteria for their identification (Calzolari *et al.* 2002; Gries 2008).

The *Dici* is basically an on-going project that is updated and incremented regularly; the following are some of the features that we plan to add shortly:

- the possibility of handling deeper syntactic modifications of word combinations (passives, relatives, left and right dislocations, etc.);
- a stronger support for writing activities, with the possibility to receive automatic suggestions with a list of existing combinations for a given word (Chang *et al.* 2008);
- the distribution of the collocations included in the dictionary within the different levels of competence of the *Common European Framework* (Council of Europe 2001).

References

ALDABE, I., ARRIETA, B., DÍAZ DE ILARAZZA, A., MARITXALAR, M., NIEBLA, I., ORONOZ, M. and URIA, L. (2006). The use of NLP tools for Basque in a multiple user CALL environment and its feedback. In P. Mertens, C. Fairon, A. Dister and P. Watrin (eds). *Verbum ex machina. Actes de la 13e conférence sur le Traitement automatique des langues naturelles*. Louvain-la-Neuve: Presses universitaires de Louvain: 815-824 (*Cahiers du Cental* 2).

ATZORI, L., CHIAPEDI, N. AND SPINA S. (2009). Corpora di italiano L2: difficoltà di annotazione e trascrizione "allargata". In C. Andorno and S. Rastelli (eds). *Corpora di italiano L2: tecnologie, metodi, spunti teorici*. Perugia: Guerra: 93-110.

ALONSO RAMOS, M. (2003). Hacia un Diccionario de colocaciones del español y su codificación. In M.A. Martí *et al.* (eds). *Lexicografía computacional y semántica*. Barcelona: Universitat de Barcelona: 11-34.

BENSON, M. (1990). Collocations and general-purpose dictionaries. *International Journal of Lexicography*, 3(1): 23-35.

BORTOLINI, U., TAGLIAVINI, C. and ZAMPOLLI, A. (1971). *Lessico di frequenza della lingua italiana contemporanea*. Milano: Garzanti.

CALZOLARI, N., FILLMORE, C., GRISHMAN, P., IDE, N., LENCI, A., MACLEOD, C. and ZAMPOLLI, A. (2002). Towards Best Practice for Multiword Expressions in Computational Lexicons. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands; Spain, 29 May – 31 May 2002: 1934-1940.

CHANG, Y., CHANG, J. S., CHEN, H. and LIOU, H. (2008). An Automatic Collocation Writing Assistant for Taiwanese EFL Learners: A Case of Corpus-Based NLP Technology. *Computer Assisted Language Learning*, 21 (3): 283-299.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

COWIE, A. (1981). The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics*, 2: 223-235.

DE MAURO, T., MANCINI, F., VEDOVELLI, M. and VOGHERA, M. (1993). *Lessico di frequenza dell'italiano parlato*. Milano: Etas.

EVERT, S. (2005). *The statistics of word cooccurrences: Word pairs and collocations*. Stuttgart: IMS.

GRANGER, S. (2004). Computer learner corpus research: current status and future prospects. In U. Connor and T. Upton (eds). *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam & Atlanta: Rodopi: 123-145.

GRANGER, S. and PAQUOT, M. (2008). Disentangling the phraseological web. In S. Granger and F. Meunier (eds). *Phraseology. An interdisciplinary perspective*. Amsterdam: John Benjamins: 27-49.

GRIES, S. (2008). Phraseology and linguistic theory. In S. Granger and F. Meunier (eds). *Phraseology. An interdisciplinary perspective.* Amsterdam: John Benjamins: 3-25.

HEID, U. (2008). Computational phraseology. In S. Granger and F. Meunier (eds). *Phraseology. An interdisciplinary perspective.* Amsterdam: John Benjamins: 337-360.

JAÉN, M.M. (2007). A Corpus-driven Design of a Test for Assessing the ESL Collocational Competence of University Students. *International Journal of English Studies*, 7(2): 127-147.

JEŽEK, E. (2005). *Lessico: classi di parole, strutture, combinazioni*, Bologna: Il Mulino.

JONES, S. and SINCLAIR, J. (1974). English lexical collocations. *Cahiers de Lexicologie*, 24: 15-61.

JUILLAND, A. and TRAVERSA, V. (1973). *Frequency dictionary of Italian words.* The Hague: Mouton.

KENNEDY, G. (2008). Phraseology and language pedagogy. In F. Meunier and S. Granger (eds). *Phraseology in Foreign Language Learning and Teaching.* Amsterdam: John Benjamins: 21-41.

LEWIS, M. (2000). *Teaching collocation. Further developments in the lexical approach*. Hove: Language Teaching Publications.

MEUNIER, F. and GRANGER, S. (eds) (2008). *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins.

NESSELHAUF, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.

NATION, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

PAZOS BRETAÑA, M. and PAMIES BERTRÁN, A. (2008). Combined statistical and grammatical criteria. In S. Granger and F. Meunier (eds). *Phraseology. An interdisciplinary perspective*. Amsterdam: John Benjamins: 391-406.

ROVENTINI, A., ALONGE, A., CALZOLARI, N., MAGNINI, B. and BERTAGNA, F. (2000). ItalWordNet: a Large Semantic Database for Italian. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation* (LREC 2000, Athens, Greece, 31 May – 2 June 2000). Paris: The European Language Resources Association (ELRA): 783-790.

SAG, I.A., BALDWIN, T., BOND, F., COPESTAKE, A., and FLICKINGER, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics* (CICLING 2002). Mexico City: 1-15.

SANTOS PEREIRA, L.A. and MENDES, A. (2002). An electronic dictionary of collocations for European Portuguese: methodology, results and applications. In A. Braasch and C. Povlsen (eds.). *Proceedings of the Tenth EURALEX International Congress* (EURALEX 2002). Copenhagen: Center for Sprogteknologi: 841-849.

SCHMID, H. (1994). Probabilistic part-of-speech tagging us- ing decision trees. *Proceedings of the International Conference on New Methods in Language Processing* (http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf).

SINCLAIR, J. (1991) *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.

SINCLAIR, J., DALEY, R., JONES, S. and KRISHNAMURTHY, R. (2004). *English collocation studies: the OSTI report*. London-New York: Continuum International Publishing.

SMADJA, F. (1993). Retrieving collocations form text: Xtract. *Computational Linguistics*, 19(1): 143-177.

SPINA, S. (2009). Building a suite of online resources to support academic vocabulary learning. In *EUROCALL 2009 – New Trends in CALL: Working Together* (Gandia, Spain, 9-12 september).

SUNG, J. (2003). *English lexical collocations and their relation to spoken fluency of adult non-native speakers*. Unpublished doctoral dissertation, Indiana University of Pennsylvania, Pennsylvania.

TSCHICHOLD, C. (2006). Intelligent CALL: The magnitude of the task. In P. Mertens, C. Fairon, A. Dister and P. Watrin (eds). *Verbum ex machina. Actes de la 13^{e} conférence sur le Traitement automatique des langues naturelles*. Louvain-la-Neuve: Presses universitaires de Louvain: 806-814 (*Cahiers du Cental* 2).

TSCHICHOLD, C. (2008). A computational lexicography approach to phraseologisms. In S. Granger and F. Meunier (eds). *Phraseology. An interdisciplinary perspective*. Amsterdam: John Benjamins: 361-376.

An ontology-based approach to the multilingual complexity of law

Daniela Tiscornia
Consiglio Nazionale delle Ricerche,
Istituto di Teoria e Tecniche per l'Informazione giuridica, ITTIG-CNR

Abstract

The need to extract relevant meaning from legal texts is an emerging issue in Legal Information handling, as well as within the NLP community, which shows an increasing interest in testing computational linguistic solutions in the field of law. In this scenario, the importance of semantic resources in the development of methodologies and tools for legal knowledge representation has been clearly assessed. The representation of legal concepts in ontological frameworks helps the understanding, sharing, and re-use of knowledge bases, as it makes explicit assumptions about legal knowledge, mediates storage of legal rules and supports reasoning on them. The aim of this article is to set out methodological routes for constructing legal ontologies in applications that, due to the tasks they intend to achieve, should maintain a clear reference to the source texts. Considering the complexity of the legal domain, methodologies based on ontology learning techniques seem to be the most promising way to fill the gap between dogmatic conceptual models and lexical patterns extracted from texts by combining lexical information and formal semantics, by codifying relations between concepts, their linguistic realisations and the contexts that contain them.

Keywords: legal language, ontologies, multilingual information.

1. Introduction

The legal terminologies used in both European and non-European legal systems express not only the legal concepts which operate in the different countries, but further reflect the deep differences existing between the various systems and the varying interpretations of lawyers in each system. Given the structural domain specificity of legal language, we cannot speak about “translating the law” to ascertain correspondences between legal terminology in various languages, since the translational correspondence of two terms satisfies neither the semantic correspondence of the concepts they denote nor the requirements of the different legal systems.

In computational applications, ontology-based models offer good solutions to cross-lingual mapping, as they allow for the representation of legal concepts in formal frameworks, thus expressing, in a coherent way, the links among the conceptual characterisation, the lexical manifestations of its components and the universes of discourse that are their proper referents. Moreover, conceptual models clarify and

explain the implicit assumptions underlining legal knowledge; in this way, understanding, sharing, and re-using knowledge bases is supported in a consistent way. Since the legal domain is strictly dependent on its own textual nature, a methodology for ontology construction must privilege a bottom-up approach, based on a solid theoretical model. Considering the complexity of the law, methodologies based on ontology learning techniques seem to be the most promising way to fill the gap between dogmatic conceptual models and lexical patterns extracted from texts.

Ontology learning techniques are tools for automatically extracting information from texts and for giving a structured organisation to such knowledge; this means combining lexical information and formal semantics, by codifying relations between concepts, their linguistic realisations and contexts. To grasp the characteristics of the legal world, such a methodology requires specific approaches able to capture the multi-layered structure of the discourse, the extensional semantics of open textured concepts and the ontological characterisation of the interpretative process in concepts construction.

The aim of this article is, therefore, to set out, through the description of two projects that have been implemented by the author, methodological routes for constructing legal ontologies in applications that, due to the tasks they intend to achieve, should maintain a clear reference to the source texts. The article is structured as follows. In section 2, we address methodological issues, by analysing the relations between language and law and the semantic relations among the levels of legal discourse. We set out models of analysis of legal documents and coherent methods for processing them in section 3. In section 4, we describe the outcomes of recent projects, focusing on the role of a middle-out conceptual level for connecting lexical and ontological layers. Finally, we comment on the lessons we have learned, outline methodological lacunae and indicate the future directions of our research.

2. Language and law

There is a strict connection between law and language, characterised by the coexistence of two autonomous but structurally similar systems: both are endowed with *rules* that underlie the construction of the system itself, guide its evolution and guarantee its consistency. Both are conditioned by the social dimension in which they are placed, whereby they dynamically define and fix their object in relation to a continually evolving social context.

The interrelation between language and law is not symmetrical because there is a strict *dependence* of the law on its linguistic expression: the law has to be communicated and social and legal rules are mainly transmitted through their oral and written expression. Even in customary law, there is almost always a phase of verbalisation that makes it possible to be identified or recognised; even if the law cannot be reduced to a language that expresses it, it cannot escape its textual nature.

Another characteristic of the law is that it is expressed through many levels of discourse:

- *legislative language* is the "object" language because it is the principal source of positive law that, in the broad sense, also includes contracts and so-called soft law*;* the constitutive force of written sources derives from the *stipulative* nature of legislative definitions and of authentic interpretations that assign a conventional meaning of legal concepts in relation to the domain covered by the law that contains them;
- *judges* interpret legal language in an "operative" sense in order to apply norms to concrete cases: the main function of judicial discourse lies, therefore, in populating the extensional dimension of the object language, instantiating cases throughout judicial subsumption, that is, linking general and abstract legislative statements to their linguistic manifestation, or, in other words, classifing legal case elements according to legal issues;
- the language of *doctrine* is a reformulation of legislative and judicial language aimed at the conceptualisation of the normative contents: whilst being a metalanguage with respect to legislative and judicial language, it is a linguistic object as well, based on the analysis of the universe of the discourse;
- *legal theory* expresses the basic concepts, the systemic categories of the legal system (for example, *subjective right*, *liability*, *sanction*, *legal act*, *cause*, *entitlement*...). The building blocks and construction rules of legal theory are independent of an observable reality, but also of positive legal systems, which constitute possible *models* of them. Their role is mainly syntactic, since they provide a systematic structure to the regulative organisation of social communities (Ross 1951). Legal theory may, therefore, be constructed as a formal and axiomatic system, made up of concepts and assertions in the theory, whose scope is explanatory of positive legal systems (Ferraioli 2007: 47);
- and finally, there is the discourse of *philosophy of law*, expressing both general principles and value judgements as well as their ordering criteria.

On the (meta)theoretical level, the border between legal theory and doctrine may be seen as a *genus/species* relationship or, in a model-theoretic interpretation, as a relation between a logical theory and its models; legal theory has an *explanatory* and *prescriptive* function - in the broad sense - because it constructs concepts independently of normative statements and interpretative operations, whilst the conceptual models of doctrine arise out of the analysis of legal texts which produce interpreted knowledge and are, therefore, not susceptible to being generalised in an axiomatic theory. One of the most obvious demonstrations of this distinction is the creative role of legal translation, halfway between term equivalence setting and concept comparison.

Transferred into the computational context, the boundary between conceptualisations of legal theory and jurisprudential models becomes purely methodological. The former entities, the *kernel legal concepts*, are modelled in the so called *core ontologies*, while the latter are the object of *domain ontologies.* Both are expressed by the same languages and by the same cognitive perspectives: computational ontologies are means of communication, aiming at making a set of meaning assumptions shared by a social community explicit; therefore, we do not claim that a domain ontology is the 'true' conceptualisation, but that it is nothing more than a partial and non-exclusive interpretation of a piece of social reality.

Since the focus of this article is to analyse legal ontologies from the perspective of linking language to concepts, in order to handle multilinguality, we will not go into details about consolidated core legal ontologies. Instead, we will concentrate on the design of a methodology that best reformulates, in a computational context, the process of jurisprudential conceptualisation, strongly based on language analysis.

3. Legal text analysis

In analysing legal documents, the basic aspect that must be elicited is the relation between the *normative statement*, which corresponds to a partition in the legal text, like articles or subsections, and the *norm*, conceived as the interpreted meaning of written regulations. A general methodology for meaning extraction must be framed through a modular architecture, where different aspects refer to specific analytic models and to appropriate NLP tools.

3.1. The semantics of textual structures

Despite the lack of specific rules governing the use of language, several legal documents have fixed narrative structures, so that it is possible to detect semantic templates from typical linguistic structures, on the basis of a finite set of linguistic expressions and syntactic structures that can be considered domain independent.

As for legislation, we can define a model of the *logical structure of legislative texts*, understood as a set of statements that enables the following elements to be identified:

- information about the document structure, for example, enacting authority, class of source, time, publishing date, versioning, subject, partitions, *etc.* (Agnoloni *et al.* 2007);
- classification of legislative statements according to their illocutive function, for example, to define, to prohibit, to oblige, to sanction (Biagioli *et al.* 2008);
- distinction of language levels, for example, norms that talk about other norms, to repeal them, to amend them, etc. (Spinosa *et al.* 2009).

The model of the *inner structure of legislative statements* is based on:

- interpretation of syntactic elements (*even if, unless, notwithstanding, and/or, but otherwise, after, etc.*) in terms of logical connectives among propositions

(Allen 1986); distinction between the *deontic classification* of behaviour and the *set of regulated behaviours*: the former is *domain independent* knowledge, the latter expresses *common sense* knowledge (Francesconi 2009).

Case law requires a different profile:

- analysis of rhetorical structures in legal judgments, to identify the basic components: facts of the case, decisions, arguments and grounds (Wyner *et al.* 2008) and of the logical argumentative structure of the case (Wyner *et al.* 2009).

3.2. The computational models

Computational frameworks, covering the whole scenario outlined above are quite difficult to find in the literature on Artificial Intelligence and Law. Traditional rule-based and case-based approaches developed in the 80s were interested in capturing the inferential aspects of legal knowledge more than in expressing the conceptual components and dependencies among different kinds of knowledge.

Proposals at the theoretical level adopted frame-based descriptive formalisms which, at present, turn out to be more easily reformulated in ontological languages. Norm frame (van Kralingen 1993) is a general structure, where frame elements – *slots* – expressing properties of legal documents are distinguished from the attributes (frame elements) of norm components. The Functional Ontology of Law (Valente 1995; Breuker *et al.* 2004), even though presented as a core ontology, is more addressed towards describing the epistemological aspects of law as a control system of social behaviours.

These models are embedded in *core legal ontologies*, usually built *top-down* with the goal of representing intensional descriptions of legal concepts as classes for guiding the interpretation of the world and explaining common sense reasoning. Formal ontologies are composed of a relatively small set of concepts, defined by a high number of constraints which encode the relations between individuals of classes through cardinality restrictions, property range and domain, disjointness, transitive and symmetric properties. The LKIF Core LKIF-Core ontology,[1] developed within the ESTRELLA project is a modular collection of basic legal concepts aimed at supporting the implementation of rule-based knowledge bases for regulatory decision support systems (Breuker *et al.* 2007). The Core Legal Ontology (CLO) (Gangemi *et al.* 2003) organises legal concepts and relations on the basis of formal properties defined in the DOLCE+ foundational ontology library (Masolo *et al.* 2002).[2]

Core ontologies are normally built on the knowledge elicited from legal experts and include the formalisation of basic concepts with which legal theory commonly agrees. In their specialisations in domain ontologies, the choice about the levels of

[1] http://www.estrellaproject.org/lkif-core (Breuker *et al.* 2007).
[2] http://dolce.semanticweb.org

generalisation is left to the developers; it mainly depends on the kind of applications and the results one expects to achieve, as they are expected to support classification, reasoning and the decision making process.

4. A bottom up methodology for ontology building

While several knowledge elements listed in section 3.1. can be semi-automatically recognized and extracted by means of machine learning and text mining tools, we cannot expect that the automatic acquisition of elements of knowledge is able to capture the complete meaning of normative statements, such as roles, legal effects, temporal and spatial parameters. Therefore, ontology construction still requires large amounts of knowledge to be manually drafted. The best solution in practical applications (*e.g.*, cross-lingual mapping) seems to be a *middle out* methodology, where text processing techniques and knowledge formalisation methods interact, through an iterative process of ontology learning and enrichment.

Many currently available Natural Language Processing (NLP) applications are rapidly evolving from the traditional processing of the formal aspects of language (part of speech tagging, syntactic parsing) towards automated analysis of meaning, by implementing tools for "ontology learning" (the term denotes a suite of methodologies and procedures for extracting the semantic structure from linguistic objects). The parsing process works in layers of increasing complexity, exemplified in the so called *ontology learning cake* (Gomez-Perz and Manzano-Macho 2003) and explained in Buitelaar *et al.* (2006: 10): ontology development is primarily concerned with the definition of concepts and relations between them, but connected to this also knowledge about the symbols that are used to refer to them. In our case this implies the acquisition of linguistic knowledge about the terms that are used to refer to a specific concept in the text and possible synonyms of these terms. An ontology further consists of a taxonomy backbone and other, non-hierarchical relations. Finally, in order to derive also facts that are not explicitly encoded by the ontology but could be derived from it, also rules should be defined (and if possible acquired) that allow for such derivations.

4.1. Lexical ontologies

At the lower layer, terms are extracted and organised in *semantic lexicons*. A de facto standard for building lexical ontologies is WordNet (Fellbaum 1998), a lexical database which has been under constant development at Princeton University. It has been described by Felbaum and Vossen (2008) as follows:

> "In fact WordNet merely attempts to map the lexicon into a network organized by means of relations... A lexicon can be defined as the mappings of words onto concepts."

Modelled along the same lines as the Princeton WordNet[3], the EuroWordNet (EWN) project is a multilingual lexical database with wordnets for eight European languages (Vossen *et al.* 1997).

Our experience in building a multilingual wordnet in the legal domain relates to the realisation of the LOIS (Lexical Ontologies for Legal Information Sharing) database,[4] composed of about 35,000 concepts in five European languages (English, German, Portuguese, Czech, and Italian) linked by English (Peters *et al.* 2007). In LOIS, a concept is expressed by a *synset*, the atomic unit of the semantic net. A synset is a set of one or more uninflected word forms belonging to the same part-of-speech (*noun, verb, adjective*) that can be interchanged in a certain context. For example, {*action, trial, proceedings, law suit*} form a noun-synset because they can be used to refer to the same concept. More precisely, each synset is a set of *word-senses*, since polysemous terms are distinct in different word-senses, *e.g.*, {*diritto_1(right)*}and {*diritto_2(law)*};{*property_1, attribute, dimension*} and {*property_2, belongings, holding*}. Each word sense belongs to exactly one synset and each word sense has exactly one word that represents it lexically, and one word can be related to one or more word senses. A synset is often further described by a gloss, explaining the meaning of the concept.

In monolingual lexicons, *word senses* are linked by lexical relations: *synonym* (included in the notion of synset), *near-synonym*, *antonym*, *derivation*. Synsets are structured by means of hierarchical relations (*hypernymy/hyponymy*) and non hierarchical relations of which the most important are *meronymy* (between parts or wholes), *thematic roles*, and *instance-of*.

In building the LOIS database, we had to choose between two basic approaches for automatic multilingual alignment of wordnets (Vossen 1999): the *extend* approach, from the source lexicon to target wordnets by means of term-to-term translation; and the *merge* approach, according to which an independent wordnet for a certain language is first created and then mapped to the others. This approach, adopted in the EuroWordnet project, determines the interconnectivity of the indigenous wordnets by means of the Inter-Lingual-Index (ILI), a set of equivalence relations of each synset with an English synset. Cross-lingual linking indicates complete equivalence, near-equivalence, or equivalence-as-a-hyponym or hyperonym. Unlike the wordnets, the ILI is a flat list and, unlike an ontology, it is not structured by means of relations. ILI entries merely function to connect equivalent words and synsets in different languages.

This solution, adopted in LOIS, seems coherent with the assumption that, in legal language, every term collection belonging to a language system, and any vocabulary originated by a law system is an autonomous lexicon and should be mapped through equivalence relationships to a pivot language, which acts as the interlingua. Language-specific synsets from different languages linked to the same ILI-record by means of a

[3] wordnet.princeton.edu and www.globalwordnet.org
[4] LOIS, Lexical Ontologies for Legal Information Sharing (EDC 2026-2001-02.)

synonym relation are considered conceptually equivalent. The LOIS database has been built in a semi-automatic way, by means of NLP techniques for morpho-syntactic parsing and conceptual clustering, to extract syntagmatic and paradigmatic relations between terms; from the output, sets of candidates for synonyms, taxonomies and non-hierarchical relations were further manually refined.

Computational lexicons, even though sometimes called *lightweight ontologies*, are in fact able to capture only the lexical semantic of terms, whose meaning merely depends on their position in the network and on semantic relations.[5] The definition of cross-lingual equivalence relations is supported by the English gloss and it is, therefore, language dependent. Nevertheless, computational lexicons provide powerful support for semantic classification, cross-lingual retrieval, term extraction, semantic interoperability, etc.[6]

In the legal domain, the shallow semantic characterization of synsets may generate ambiguities or loss of information in cases where domain and linguistic information overlap. To give some examples, in LOIS, sub-class relations (*i.e., rental contract, contract*), and semantic specialisation *(i.e., unfair competition, competition*) are not distinct; the semantic notion of functional equivalence is adopted to express the legal notion of "similarity in functions" (*i.e. Camera dei Deputati, Assemblée nationale, Congreso de los Diputados*). Most of the semantic combinatorial properties of lexical items are not explicitly represented and multi-words like "*place of contract conclusion", "offer acceptance", "contract infringement"* cannot be expressed. Any new legislative definition is considered to be a conventional assignment of a new meaning to a concept, and, consequently, the introduction of a new sense (Tiscornia 2006). The lesson learned from the LOIS experience brought us to the conclusion that semantic soundness and consistent integration of semantic lexicons must be supported by a domain-specific ontology, in order to make explicit the way in which legal concepts belonging to different systems share analogies and set differences.

4.2. Anchoring terminologies to a reference ontology

Our experience of interfacing lexicons and ontologies on the DALOS (DrAfting Legislation with Ontology-based Support[7]) project (Agnoloni *et al.* 2007) arose out of the task the project intended to perform, *i.e.*, the creation of a terminology control tool in the multilingual drafting of Community legislation, where the institutional texts

[5] According to Hirst (2004), "A lexicon is not a very good ontology. An ontology, after all, is a set of categories of objects or ideas in the world, along with certain relationships among them; it is not a linguistic object. A lexicon, on the other hand, depends, by definition, on a natural language and the word senses in it. [...] Despite all the discussion in the previous section, it is possible that a lexicon with a semantic hierarchy might serve as the basis for a useful ontology, and an ontology may serve as a grounding for a lexicon. This may be so in particular in technical domains, in which vocabulary and ontology are more closely tied than in more-general domains."

[6] An OWL meta ontology for connecting Wordnets is available at: http://www.w3.org/TR/wordnet-rdf/

[7] www.dalosproject.eu

expressed in the 25 languages of the European Union are deemed semantically and, therefore, normatively equivalent.

The aim, therefore, of DALOS is to provide a knowledge and linguistic resource for legislative drafting. The main outcome is the definition of a semantic framework, where the use of words and the underlining meaning assumptions are made explicit. Such knowledge, embedded in a specialised drafting tool, will provide law-makers with a clear overview of the consolidated lexicon in a regulative domain and of the semantic properties of concepts, thus facilitating the harmonisation of legal knowledge and lexicons between the EU and Member States. It also supports the dynamic integration of the lexicon by the legislator and the monitoring of the diachronic meaning evolution of legal terminology. The knowledge base, as shown in Figure 1, is composed of:

1. the Lexical Layer which contains lexicons extracted by means of NLP tools[8] from a set of parallel corpora of EU legislation and case law in the sub-domain of consumer protection, chosen as a case study (16 EU Directives, 33 European Court of Justice decisions and 9 Court of First Instance judgments).[9] Extracted terminologies[10] have been manually refined, producing four monolingual terminologies (in Italian, English, Dutch and Spanish), structured along the lines of WordNet, and formally codified as sets of instances of the Noun-Synset class, identified by a Uniform Resource Identification and described by OWL object properties that translate WordNet relations. Each word sense is also linked to its textual referent, a text fragment codified as an instance of class partition (Agnoloni *et al.* 2009).

[8] The tools, specifically addressed to process English and other EU language texts, are GATE and T2K. GATE supports advanced language analysis; Gate is distributed and maintained by the Department of Computer Science of the University of Sheffield. T2K is a terminology extractor and ontology learning tool for the Italian language jointly developed by CNR-ILC and the University of Pisa.

[9] In order to guarantee the linking of acquired domain terms to the individual textual partitions rather than to the individual act, the corpus to be processed was segmented into 8,192 files corresponding to 2,583 directive partitions (sub-paragraphs) and 5,609 case law partitions.

[10] The selected minimum frequency threshold for both single and multiword terms was 5. The percentage of selected terms from the ranked lists was 20% and 70% for multiword terms. The Italian TermBank is composed of 1,443 terms of which 1,168 are multiword terms of different complexity. The number of extracted hyponymic relations is 623 referring to 229 hypernym terms, whereas the number of identified related terms is 1,258 referring to 279 terminological headwords. The processing of the English corpus resulted in a set of 3,012 terms, which consists of 1,157 multi-word units and 1,855 single word terms. This set has an overlap of 572 terms with the LOIS vocabulary. Crosslingual alignment computes the overlap between the different languages according to two criteria: a) the positional similarity in the texts; b) (near)equivalents on the basis of translations (through WordNet). See Agnoloni *et al*. (2009) for more detail.

2. the Ontological Layer built on top of the lexicon is composed of:

- the *Concept Layer*, a flat list of synsets, linked by a "has-lexicalisation" relation to monolingual synsets in the lexical layer; it acts as a centralized index, like the Interlingual Index (ILI) in Wordnet, composed of all the equivalence relations of monolingual wordnets to the English Wordnet, in order to align synsets of different languages. The index provides the extensional characterization of concepts without carrying any kind of semantic information, which is provided by the ontology;
- the *Domain Ontology* which formally describes the intensional meaning of core elements in the consumer law domain. In selecting candidates for the ontology, we assumed that all concepts defined in the legislative corpus are relevant, as well as several concepts used in the definitional contexts, expressing the basic properties of the domain, which has been modelled around the notion of *commercial transaction*, relying on the basic state of affairs regulated. The role of the ontological layer is to assign a domain specific characterization to entities at the conceptual level, and consequently, to 'explain' and validate terminological choices at the lexical layer.

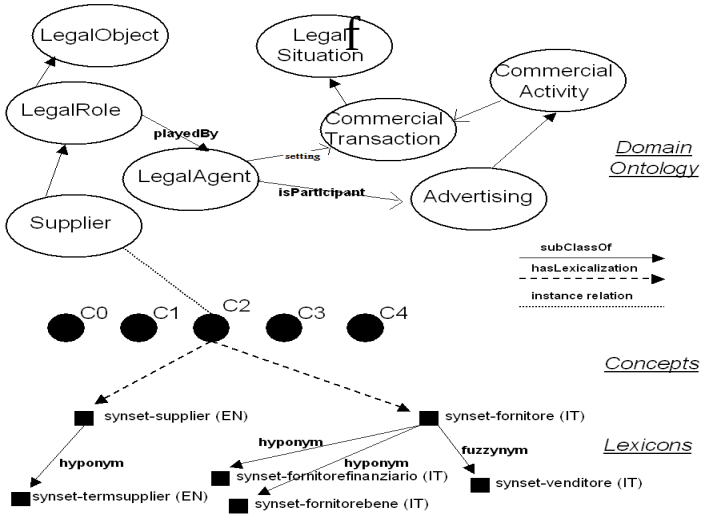


Figure 1. Knowledge base architecture

Through the double-layered representation, concept expressions (*provider, producer, importer; produttore, fornitore, distributore, etc.*) extracted from documents are 'assigned' to the same conceptual entity 'supplier', whose basic properties are formally codified in the ontology. In this way, meaning constraints imposed by

legislative definitions can be anchored to their normative scope and misleading alignments or diverging lexicalisations can be detected and explained.

5. Conclusion

The article has discussed some issues related to the construction of multilingual legal lexicons. Beginning from the illustration of analytical profiles of legal texts, it describes the ontology-based approach, a methodology that seems more reliable than traditional approaches to terminology building, as it is context-sensitive and, therefore, similar to the processes for concept formation elaborated by legal doctrine.

In this scenario, the importance of solid semantic resources in the development of tools for cross-lingual linking has been assessed, and, from lessons learned in practical applications, the potential of ontology learning techniques, as a powerful approach to legal concept modelling, compliant with the contextual nature of legal semantics, has been described.

The project presented here is expected to be a further step towards the ambitious goal of automatizing the passage from lexical to formal knowledge, by using tools and procedures that enable the two levels to be brought closer together, so diminishing both the burden of the manual work and the arbitrary nature of the links.

References

AGNOLONI, T. FRANCESCONI, E. and SPINOSA, P. (2007). XmlLegesEditor: an OpenSource Visual XML Editor for supporting Legal National Standards. In C. Biagioli, E. Francesconi and G. Sartor (eds). *Proceedings of the V Legislative XML Workshop*. Florence: European Press Academic Publishing: 239-251.

AGNOLONI, T., BACCI, L., FRANCESCONI, E., PETERS, W., MONTEMAGNI, S. and VENTURI, G. (2009). A two-level knowledge approach to support multilingual legislative drafting. In J. Breuker, P. Casanovas, E. Francesconi and M. Klein (eds). *Law, Ontologies and the Semantic Web,*.Amsterdam: IOS Press.

ALLEN, L.E. and SAXON, C. (1996). Analysis of the Logical Structure of Legal Rules by a Modernized and Formalized Version of Hohfeld's Fundamental Legal conceptions. In A.A. Martino and F. Socci (eds.), *Automated Analysis of Legal Texts: Logic, Informatics and Law*. Amsterdam: North-Holland: 385-450.

BUITELAAR, P., CIMIANO, P. and MAGNINI, B. (eds). 2006. *Ontology learning.* Amsterdam: IOS Press.

BIAGIOLI, C. and GROSSI, D. (2008). Formal Aspects of Legislative Meta-drafting. In E. Francesconi, G. Sartor and D.Tiscornia (eds). *Legal Knowledge and Information Systems - JURIX 2008*: The Twenty-First Annual Conference. Amsterdam: IOS Press: 192-201.

BREUKER, J., VALENTE, A. and WINKELS, R. (2004). Legal Ontologies in Knowledge Engineering and Information Management. *Artificial Intelligence and Law*, Kluwer, 12(4): 241-277.

BREUKER, J., HOEKSTRA, R., BOER, A., VAN DEN BERG, K.,RUBINO, R., SARTOR, G., PALMIRANI, M., WYNER, A. and BENCH-CAPON, T. (2007). *OWL ontology of basic legal concepts (LKIF-Core).* Deliverable 1.4, Estrella.

FELLBAUM, C. (ed.) (1998). *WordNet: An electronic lexical database*. Boston: MIT Press.

FERRAIOLI, L. (2007). *Teoria del diritto e della Democrazia*, Volume primo. Bari: Laterza.

FRANCESCONI, E. (2009). An Approach to Legal Rules Modelling and Automatic Learning. In *Proceedings of Jurix 2009 Conference*. Amsterdam: Ios Press.

FELBAUM, C, and VOSSEN, P. (2008). *Challenges for a global WordNet.* In *Proceedings of The First International Conference on Global Interoperability for Language Resources*. Hong Kong: 75-81.

GANGEMI, A., SAGRI, M.T. and TISCORNIA, D. (2003). A Constructive Framework for Legal Ontologies. In R. Bejamins., P. Casanovas, J. Breuker and A. Gangemi (eds). *Law and the Semantic Web: Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications.* Amsterdam: IOS Press: 97-124.

GOMEZ-PERZ, A. and MANZANO-MACHO, D. (2003). *A survey of ontology learning methods and techniques.* Ontoweb Deliverable 1.5.2003.

HIRST, G. (2004). Ontology and the Lexicon. In S. Staab and R. Studer (eds.). *Handbook on Ontologies in Information Systems.* Berlin: Springer: 209-230.

MASOLO, C., BORGO, S., GANGEMI, A., GUARINO, N., OLTRAMARI, A. and SCHNEIDER, L. (2002). *The WonderWeb Library of Foundational Ontologies and the DOLCE ontology.* WonderWeb Deliverable D17.

MASOLO, C., VIEU, L., BOTTAZZI, E., CATENACCI, C., FERRARIO, R., GANGEMI, A. and GUARINO, N. (2004). Social roles and their descriptions. In C. Welty (ed.). *Proceedings of the Ninth International Conference on the Principles of Knowledge Representation and Reasoning.* Whistler.

PETERS, W., TISCORNIA, D. and SAGRI, M.T. (2007). The Structuring of Legal Knowledge in Lois. In *Artificial Intelligence and Law.* Special Issue on Legal knowledge extraction and searching and legal ontology applications.15.2:117-135.

ROSS, A. (1951). *"Tû-Tû". Harvard law review* 70 (5): 812-825. Originally published in *Festskrift til Henry Ussing*. O. Borum, K. Ilium (eds). Kobenhavn Juristforbundet, 1951.

SPINOSA, P., CHERUBINI, M., GIARDIELLO, G., MARCHI, S., MONTEMAGNI, S., and VENTURI, G. (2009). *Legal Texts Consolidation through NLP-based Metada Extraction.* In Proceedings of ICAIL 2009, Barcelona, ACM Press.

VALENTE, A. (1995). *A functional ontology of law.* Amsterdam, Information Science Institute.

VAN KRALINGEN R. (1993). A Conceptual Frame-based Ontology for the Law. In *Proceedings of Jurix 1993*. Amsterdam: IOS Press.

TISCORNIA, D. (2006). The Lois Project: Lexical Ontologies for Legal Information Sharing. In C. Biagioli, E. Francesconi and G.Sartor (eds). *Proceedings of the V Legislative XML Workshop.* Florence: European Press Academic Publishing.

VOSSEN P. (ed.) (1998). *EuroWordNet: a multilingual database with lexical semantic networks.* Berlin: Springer.

VOSSEN P., PETERS W. and GONZALO J. (1999). Towards a Universal Index of Meaning. In *Proceedings of ACL-99* Workshop. Berlin: Springer, Lecture Notes in Computer Science Subseries.

WYNER A., MOCHALES-PALAU R., MOENS M.F. and MILWARD D. (2009). Approaches to Text Mining Arguments from Legal Cases. In E. Francesconi, S. Montemagni,W. Peters and D. Tiscornia (eds). *Semantic processing of legal texts.* Berlin: Springer.

WYNER, A, BENCH CAPON, T. J.M. and ATKINSON, K. (2008). Three Senses of "Argument". In P. Casanovas, G. Sartor, N. Casellas and R. Rubino (eds). *Computable models of the Law. Berlin:* Springer, Lecture Notes in Computer Science Subseries: 146-162.

Dynamic access to a static dictionary: a lexicographical "cathedral" lives to see the twenty-first century – the *Dictionnaire étymologique de l'ancien français*

Sabine Tittel
Heidelberg Academy of Sciences and Humanities

Abstract

How to transform a dictionary of the Old French language – an ongoing composition for over 40 years – with a complex and dense article structure? The dictionary's structure is characterised by rules for condensing information, explicitly by the use of conventional indicators and implicitly by the use of rules relying on the user's knowledge of the article structure. The target format was a dictionary which is printed and published online including versatile search functions which allow access both to explicit and implicit information. The transformation process clearly showed the importance of a well directed cooperation between lexicography and computer science controlled by the lexicographers.

Keywords: etymological dictionary of old French, historical language, dictionary writing system, cooperation between lexicography and computer science.

1. Introduction

New lexicographical projects start with certain advantages: computer scientists do not need to deal with the problem of how to dress a weighty tome – a long-established dictionary – in fancy new clothes. Rather, thanks to the sartorial elegance of modern computer science, the adolescent body of a newly-designed dictionary will come into the world seamlessly and gracefully formed, and able to wear anything that the modern user's requirements hope to see it sport.

To dress anew a bulky old dictionary – and to avoid the emperor's new clothes in doing so – was the task of the editorial team of the *Dictionnaire étymologique de l'ançien français* (DEAF), the Etymological Dictionary of Old French, compiled in Germany at Heidelberg University under the aegis of the Heidelberg Academy of Sciences and Humanities.

We need to visualize the substantial corpus of this dictionary as a scientifically very stable, but static and weighty entity: the dictionary's concepts, the plan of its entries, the appearance of the published product and the scientific approach were all well established.

The work-flow by which the publication's rhythm was driven was excellent, its high standing within the international scientific community unquestioned. But what was missing was a response to the questions asked by on-line users, the key challenge of twenty-first century lexicography. In other words: what was missing was the transformation of lexicography into eLexicography, a process which implied new dictionary architecture and the successful bringing together of both new elements and those which have been scientifically established for decades.

Thus the challenge we report on was how to create dynamic access to a dictionary whose own content-related structure would continue to be static. The challenge did not lie primarily in the detail of technical implementation but rather in the interface between lexicography and computer science.

The transformation necessitated firstly an entirely new interior and secondly an entirely new exterior of the dictionary.

2. The *Dictionnaire étymologique de l'ancien français*

2.1. Some facts about the DEAF

Some facts about the Dictionnaire étymologique de l'ancien français (DEAF), its source material and its corpus shall illustrate the starting point of the transformation process.

The DEAF is a historical dictionary of the Old French language which takes into account the further development of Old French to Modern French as well as all vernacular languages of medieval Europe including Mediaeval Latin. It is an etymological dictionary which involves Latin, Greek, Germanic languages, Hebrew, Arabic, etc. It has a complex and dense dictionary entry structure due to its scientific content: the information is presented in a maximally space-saving manner whilst preserving a maximal quality. The project has been running for over 40 years (volumes G to K [1974-2008]: 10,217 entries, 4,099 columns, 636 pages of printed indices). The dictionary's critical and annotated bibliography DEAFBibl (Möhren 2007) is the most comprehensive of its kind with references to more than 6,000 sources on 1,031 columns; it is the industry standard work used also by other dictionaries, journals, etc. [ANDEI[1], DMF, CCFM[2], Nouveau Corpus d'Amsterdam (Stein *et al.* 2006)]. The dictionary is based on a vast quantity of Old French sources;

[1] In this paper, we use the siglum system of DEAFBiblEl to refer to other dictionaries. DEAFBiblEl is freely accessible via http://www.deaf-page.de; to provide complete clarity, the DEAF-sigla used in this paper are included in the bibliographical references at the end of the paper.

[2] "Consortium pour les corpus de français médiéval" CCFM, founded in October 2004 by the University of Ottawa, the "École Normale Supérieure Lettres et Sciences humaines", Lyon, the University of Stuttgart, the University of Zürich, the "Laboratoire ATILF (Analyse et traitement informatique de la langue française) ", Nancy, the University of Wales, Aberystwyth, and the "École nationale des chartes", Paris. *Cf.* http://ccfm.ens-lsh.fr.

the corpus is open, including all available editions of all texts, manuscripts, secondary literature, dictionaries, etc. 1.5 million handwritten slips (fiches) with heterogeneous information leading to 12 million attestations serve as an entry into the sources (regarding the 'ad fontes' principle, *i.e.* always go back to the sources).

2.2. The DEAF's complexity

The DEAF is substantially more than a dictionary, being in many respects more a compilation of monographs on every Old French word. The article structure is due to the potential multiplicity of the individual monographs: the information given is of considerable complexity, nonetheless rendered in a clear and consistent structure.[3]

The DEAF is characterised by a very complex article structure. The article is composed of a main entry with subentries which consist of: (1) an etymological discussion (including information on the etymon's persistence in other (Romance) languages [with indication of source, part of speech, definition and dating] and various remarks); (2) graphical variants (with Old French 'scriptae', attestations which may differ from those of the semantic article section, text and manuscript dating, inflection related to the Old French two-case system); and (3) the semantic section (divided into numerous main senses and sub-senses with definitions, dating, indication of part of speech, of terminology, of additional information on the usage [*i.e.* extension, euphemism, etc.], together with cross-references to other senses, with text references giving one or several informative contexts including variant manuscript readings and comments on content, references to other dictionary entries and secondary literature, etc.). Parenthetical comments and footnotes may occur anywhere and everywhere.

Every given attestation is traced 'ad fontes' but not necessarily quoted in the article, considering the number of attestations per sense ranging from one to more than 1,000. As a rule, each sense refers to the first three textual references to each of the first ten source texts in chronological order. More than three textual references to one source text are condensed as "etc.", more than ten source texts as "etc., etc.". Relevant references to additional attestations and texts break this rule.

Conventional abbreviations / indicators are used to condense information given in the article (concerning the definitions of words in another language as well as textual references, definitions of sub-senses, indication of Old French scriptae, parts of speech, etc.). This principle of using condensed information also works in part without explicit abbreviations / indicators, relying purely on the user's knowledge of the article structure.

[3] A point of discussion ignored in this paper is the new article concept that the DEAF had to design to meet changing monetary conditions affecting the duration of the project. It is a twofold concept which consists of extensive articles of the scientifically acknowledged lexicographical quality (referred to as "DEAFplus") and of compendious articles which present the entirety of the dictionary's raw material, semantically and orthographically pre-structured (referred to as "DEAFpré"). Discussed in this paper is exclusively DEAFplus.

A comparison between the DEAF and other lexicographical works reveals that particular structural characteristics do also appear in comparable dictionaries. But, as far as we know at present, none of these show the sheer number of the DEAF's characteristics.

3. The new interior: an editorial system

The new interior is represented by a complex editorial system which allows for a technically supported and time-saving production of dictionary entries, via a system which has been exclusively designed for the use of the editorial team.

The technical solution was developed in cooperation with the "Institute for Program Structures and Data Organization" IPD under the direction of Prof. Dr. Dr. h.c. Peter C. Lockemann, University of Karlsruhe, Germany. It implies a MySQL-Database, WicketFrameWork as a user interface, Hybernate, Databinder, etc. The system combines information management (entry, slip, and bibliography data), process management (editorial work-flow), context-dependent semantic support in editing, search, sorting, and export function for data, user administration, etc. A lemmatisation-tool (Java) has been developed with regards to the spelling variation in Old French; it is based on 120 phonetic rules considering the diatopic variation in Old French and its historical development from Latin.

One of the main questions posed during the development of the editorial system was whether it would prove possible to combine technical demands, themselves inherently inflexible, with the academic freedom which the humanities always require. The answer had to be positive and was the sine qua non, the necessary condition for the project itself: for that which must be, can be (freely adapted from Palmstroem's conclusion in *The impossible fact* by Christian Morgenstern). To achieve this, the editorial team agreed on a compromise which resulted in the combination of the benefits of two ostensibly contradictory elements: 1) benefits of electronic work-flow support with fixed structures and automated data management, together with 2) the benefits of free-text editorial input, with the facility of semantic markup.

Implicit, too, in preserving the academic freedom was the question of whether it is technically possible to reproduce precisely the traditional editing procedure of the entries. The answer, given that computer science is basically able to develop solutions for virtually any suitable problem, is again positive, that is, if computer science follows the route predefined by the lexicographical methodology in question. This already hints at the final discussion as to whether eLexicography is determined by technical potential or by lexicography itself, to which we return below.

4. The new exterior: DEAF électronique

The on-line version of the dictionary called "DEAF électronique" (DEAFél) starts in 2010. It includes the publication of all entries, preserving the dictionary's static

structure (entry layout, word families, etc.) as well as a quotable version of all printed entries. It gives access to the dictionary information separately from the dictionary's structure: a versatile system of 23 combinable search functions allows for multiple queries; the conception and compatibility of search functions match the state of the art search functions already offered, for example, by the OED, the Woordenboek or the DMF 2009. The publication of the entirety of the DEAF's raw material (slips, etc.) offers research possibilities beyond the dictionary's own contribution. The on-line publication of the DEAF bibliography with its own 20 search functions is integrated into DEAFél by linking each quoted text to the corresponding bibliographical data. Hyperlinks to the ANDEl, the DMF 2009 and other on-line dictionaries build up a dictionary network.

One of the main questions concerning the new exterior was how to create dynamic access to a dictionary whose own internal structure would continue to be static.

Now, what does a static internal structure of a dictionary refer to? It refers to the DEAF's characteristic of featuring article sub-parts which are not self-sufficient and independent of the rest of the article, and thus not detachable from their context.

Some examples shall illustrate this. The first example is taken from the semantic section, which consists of main-senses and sub-senses. There are sub-senses which feature a complete definition and which thus offer independent information. But numerous sub-senses are defined using "idem" together with additional information (*e.g.* "id." used as a metaphor); these sub-senses cross-refer to the definition of a previous sense. A similar concept is used by the MED, the LEI, the TLF and by the DMF 2009. Other dictionaries exclusively define self-sufficient senses with independent information, such as the AND, the Gdf, the OED, the Goethe-Wörterbuch (Berlin-Brandenburgische Akademie der Wissenschaften, Akademie der Wissenschaften zu Göttingen, Heidelberger Akademie der Wissenschaften 1978 →) or the Woordenboek.

The second example is taken from the graphical variants' section. Graphical variants are often marked by one or by several scriptae respectively Old French 'dialects'. Without being expressed explicitly, these scripta indications apply until such time as a different scripta is indicated or until the scripta indication is explicitly cancelled.[4]

The third example is taken from the article part which contains the etymological discussion. When tracing the etymon's persistence in other (Romance) languages a word dealt with is not followed by its definition unless its sense differs from the sense given above (other Romance word or etymon). The same applies to the indication of the part of speech.

This principle of condensing information within a static internal structure implies a considerable disadvantage: the intuitive use of the dictionary only gives access to the

[4] This is expressed by "*s.l.*" meaning "sans localisation" (*sine loco*).

information given in an explicit way but not to the information given in an implicit way.

It is what we call “dynamic access”, which allows for an inclusion of this implicit information, and which requires an elaborate structural XML markup. It appears that access to implicit information of other dictionaries via search functions does not always work with the desired accuracy, a fact which seems typical to us.

To take an example from another dictionary: in its article *veine*, the MED registers the anatomical term *veine aborchi* as ~ *aborchi* referring to the entry *veine* and being the mediaeval term for what we now define as the aorta. Not knowing that, and searching the online version of the MED for the term *veine aborchi*, the search unfortunately does not generate any matches. To find the term it is necessary to search for the adjective *veine* alone.

The DEAF has set as a target that the static form of the dictionary will not affect the searchability of the information which is explicitly and implicitly given in the dictionary’s articles.

5. Conclusion

This brings us to the final question: is eLexicography determined by technical potential, that is, by “e”, or by lexicography?

Both answers are possible and both produce results. The best solution is however that both sides – lexicography and computer science – should emphasize and insist on what is most substantial from their own point of view and fuse these elements for maximum unified effectiveness.

The transformation of DEAF into DEAFél was complicated by the range of possibilities to supply, arrange, summarize or omit information. The dictionary, we would respectfully suggest, exceeds most others in its multiplicity. And it was philosophically imperative for the lexicographers concerned that completely fixed structures and entirely automated data management should not be adopted at all prices. In fact, it is this range of options which needs to be captured to insure continuous academic freedom.

There was an early moment when “e” looked likely to dominate lexicography. This almost brought the transformation process and the existence of the dictionary itself to an end, but a change of partner (for the “Institute for Program Structures and Data Organization”, *cf. supra*) saved the project.

In the case of the transformation of a long-established dictionary whose scientific stability cannot be open to discussion, it is necessarily lexicography which must dominate the “e”. In this case, lexicography must be in charge. The challenge for computer science is to add its capabilities to the transformation project in such a way as to exploit the possibilities of improvement and enrichment: for this, the necessary precondition is the existence and availability of “e”. When dressing up the bulky old

dictionary, the real test for the "e" element is to avoid inadvertently providing the emperor's new clothes.

References

Analyse et traitement informatique de la langue française ATILF – Nancy Université & Centre National de la Recherche Scientifique CNRS (2009). *Dictionnaire du Moyen Français*. Nancy. On-line version: http://www.atilf.fr/dmf [DEAF-siglum: DMF 2009].

BALDINGER, K., MÖHREN, F., STÄDTLER, Th., DÖRR, S. and TITTEL, S. (1974-2008, →). *Dictionnaire étymologique de l'ancien français (DEAF)*, vol. G-K, →. Tübingen: Max Niemeyer Verlag.

Berlin-Brandenburgische Akademie der Wissenschaften & Akademie der Wissenschaften zu Göttingen & Heidelberger Akademie der Wissenschaften (1978 →). *Goethe-Wörterbuch*. Stuttgart: Kohlhammer.

Centre National de la Recherche Scientifique CNRS (1971-1994). *Trésor de la langue française. Dictionnaire de la langue du XIX^e^ et du XX^e^ siècle (1789-1960).* Paris: Gallimard (vol. 11ss.) [DEAF-siglum: TLF]. On-line version *Trésor de la langue française informatisé*: http://atilf.atilf.fr [DEAF-siglum: TLFi].

Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften (2004). *Das Goethe-Wörterbuch im Internet*. Trier. http://germazope.uni-trier.de/Projects/GWB.

GODEFROY, F. (1880-1902). *Dictionnaire de l'ancienne langue française et de tous ses dialectes du IX^e^ au XV^e^ siècle*. Paris [DEAF-siglum: Gdf, GdfC]. http://www.classiques-garnier.com.

KURATH, H., KUHN, S.M., REIDY, J. and LEWIS, R.E. (1952 →). *Middle English dictionary*. Ann Arbor: Univ. of Michigan Press [DEAF-siglum: MED]. http://quod.lib.umich.edu/m/med.

MÖHREN, F. (2007 [1993]). *Dictionnaire étymologique de l'ancien français (DEAF). Complément bibliographique*. Tübingen: Max Niemeyer Verlag [DEAF-siglum: DEAFBibl]. http://www.deaf-page.de [DEAF-siglum: DEAFBiblEl].

MURRAY, J.A.H. (1888-1928). *A new English dictionary on historical principles.* Oxford: Clarendon [DEAF-siglum: OED]. http://www.oed.com.

PFISTER, M. and SCHWEIKARD, W. (1979 →). *Lessico etimologico italiano*. Wiesbaden: Reichert. http://germazope.uni-trier.de/Projects/WBB/woerterbuecher [DEAF-siglum: LEI].

ROTHWELL, W., STONE, L.W. and REID, T.B.W. (1977-1992). *Anglo-Norman dictionary*, London: Maney (The Mod. Humanities Research Assoc.) [DEAF-siglum: AND].

ROTHWELL, W., GREGORY, S. and TROTTER, D.A. (2005). *Anglo-Norman dictionary. Second edition*, vol. A-E, London: Maney (The Mod. Humanities Research Assoc.) [DEAF-siglum: AND²]. http://www.anglo-norman.net [DEAF-siglum: ANDEl], currently [June 2009] covering A-K in the second edition with L-Z from the first edition [1977-1992].

STEIN, A. *et al.* (2006). *Nouveau Corpus d'Amsterdam. Corpus informatique de textes littéraires d'ancien français (ca. 1150-1350), établi par Anthonij Dees (Amsterdam 1987), remanié par Achim Stein, Pierre Kunstmann et Martin-D. Gleßgen*, Stuttgart: Institut für Linguistik/Romanistik, Universität Stuttgart.

DE VRIES, M. and TE WINKEL, L.A. (1882 →). *Woordenboek der Nederlandsche Taal.* Leiden: 's Gravenhage [DEAF-siglum: Woordenboek].

Access to multiple lexical resources at a stroke: integrating dictionary, corpus and Wordnet data

Lars Trap-Jensen[1]
Society for Danish Language and Literature, Copenhagen, Denmark

Abstract

The paper presents a lexical resource, *ordnet.dk*, which brings together data from two dictionaries – both originally print dictionaries, one historical and one modern – with a contemporary reference corpus and a wordnet, all with Danish as the object language. Focus is on data exploitation across the components, dealing with onomasiological queries in the dictionary based on wordnet data, and on cross resource look-up possibilities from the three components.

Keywords: wordnet, dictionary, corpus, resource integration, Danish.

1. Project background

It is quite likely that the technological prospects of e-media will gradually change the dictionary as a genre. No compelling reason exists why dictionaries should be confined to dealing with words and their descriptions. Many e-dictionaries already include spoken pronunciation and pictures, and why not continue with encyclopaedic articles, grammar paragraphs, and translation services until, eventually, the day arrives when we simply ask our computer a question and it provides us with the appropriate answer. However, until that day we must take one step at a time, and in this paper I will look at some of the very first steps we have taken towards increased resource integration based on our experience with a Danish online dictionary site providing access to two monolingual dictionaries and a corpus of contemporary Danish.

Ordnet.dk was developed during a six-year project period that ended in 2009. The interface gives separate dictionary and corpus access but at the same time their contents are combined in various ways. Furthermore, the contents have been supplemented with wordnet data for a new function in the online version.

A few words should be said about the original resources. Both dictionaries were conceived for print publication but at different times and under different circumstances. The *Ordbog over det danske Sprog* (The Dictionary of the Danish Language, henceforth ODS), is a historical dictionary in 28 volumes which was

[1] Society for Danish Language and Literature, Christians Brygge 1, 1219 Copenhagen K, Denmark, ltj@dsl.dk

published between 1918 and 1956. Together with its five supplementary volumes, which appeared between 1992 and 2005, it covers approximately 250,000 words from the period 1700-1950. The digitization of the original manuscript was carried out as part of the current project (see *e.g.* Lorentzen and Trap-Jensen 2008), and a preliminary version has been publicly available since 2005. With regard to the current topic of integration, the ODS is, however, the least relevant of the three components.

Den Danske Ordbog (The Danish Dictionary, henceforth DDO) is where most of the integration is being explored. It is a dictionary of modern Danish covering the period following the ODS, *i.e.* from 1950 onwards. Conceived and published (2003-2005) as a print dictionary in 6 volumes, it was, however, prepared with the use of modern methods and technology. The SGML format of the data was converted to XML and considerable effort has been made to restructure the data to improve it for screen publication as part of the current project.

KorpusDK is a reference corpus of contemporary Danish. A subset of it was built as an integral part of the DDO project, this dictionary being the first corpus-based dictionary compiled for Danish. Spoken language and texts that were restricted for privacy or copyright reasons (such as private letters and diaries) were removed and new texts added when the corpus was made public under the name of *Korpus 2000* (see *e.g.* Andersen *et al.* 2000). It was mainly the name and design that were changed when it became *KorpusDK* as part of the current project but new texts have been collected on a regular basis since 2005.

DanNet was constructed on the basis of words and senses taken from the DDO in a joint work between the Society for Danish Language and Literature and the Centre for Language Technology at the University of Copenhagen (see Pedersen *et al.* 2009). Data from DanNet are used in the online version of the DDO.

It is worth noting that the data are more compatible than one might assume at first glance. The ODS and DDO were compiled at times when both theory and practice were different but even so, they were developed by the same institution, the Society for Danish Language and Literature, and within the same tradition of descriptive lexicography. There is an intimate relation between the empirical basis of the DDO and (parts of) KorpusDK, and the same is true of the DDO and the Danish wordnet, DanNet.

2. Related words in DDO

As a new feature, the online DDO offers “related words” for a substantial number of word senses. “Related words” is a thesaurus-like function which is particularly useful for language production and for (advanced) language learning purposes. It assists users in “finding the right word” when writing a text and in developing their communicative skills to express themselves creatively, with nuance and accuracy. For language learners, it provides an overview of a semantic field that is important in vocabulary

training as words are not learnt in isolation but rather in comparison to words with similar meanings.

As data from the Danish wordnet is used for this feature, a few clarifying remarks about this resource are in order.[2]

Wordnets are language technology resources that have primarily been used in information systems, for example for information retrieval, word sense disambiguation and artificial intelligence applications. The basic unit is the *synset*: a set of one or more synonymous words that express the same concept. Each word sense belongs to a particular semantic class – called *ontological type* – established by a rough division of the conceptual world into approximately 200 semantic classes based on principles known from traditional componential analysis. Examples of ontological types are *Natural+Substance* (*ice, lava, sand*), *Plant+Object+Comestible* (*avocado, carrot, tomato*), *Human+Object+Occupation* (*accountant, nurse, taxi driver*) and *UnboundedEvent+Agentive+Mental* (*reflect, analyze, think*).

Figure 1. Information on related words in The Danish Dictionary

A synset can be related to other synsets through various semantic relations, in DanNet a total of 18 are used. The most commonly used relations are: hyponymy,

[2] The account of DanNet data in DDO is based on the account given in Sørensen and Trap-Jensen (forthcoming).

hyperonymy, part-whole, antonymy, near-synonymy, used for, concerns and involved agent. The relations have been encoded for each individual sense and the coded outcome is what is used to calculate candidates for "Related words".

An example is shown in Figure 1. For obvious reasons, this and the following examples are in Danish but hopefully they are internationally understood. According to the underlying hyponymy hierarchy, related words are selected from three levels: more general words (indicated by “mere generelt” in Figure 1) from the superordinate level are taken from the level immediately above, *i.e.* the word or words serving as *genus proximum*; more specific words (“mere specifikt” in Figure 1) are taken from the level below, *i.e.* among the hyponyms; and finally, the last group contains words at the same level as the sense looked up (indicated by the heading “andre ord med “apparat” som overbegreb” in Figure 1), *i.e.* words that are co-hyponyms or sister terms.

The first group, the hyperonyms, is straightforward as there will always be a limited number of candidates in this group, in most cases just one. It is possible to have several words occurring as hyperonyms but only in the event that a) a concept is expressed by two or more synonymous words: the word *jeep* has as its hyperonym the synset consisting of *car, auto, automobile, machine, motorcar*, or b) if a word has more than one hyperonym: in DanNet the word for *roller skate* (Danish: *rulleskøjte*) has been encoded as a hyponym of both the synset *footwear, footgear* and of the synset *sporting requisite.* Accordingly, all the synonyms appear as more general words for *jeep* and *roller skate*, respectively.

More problematic are the co-hyponyms as there may be several thousands of them in the extreme cases. To help selecting the best suitable words, a score has been calculated to express the similarity between the entry word and the co-hyponyms. The entry word in the relevant sense is compared to each of the co-hyponyms, and first the ontological types are considered: the greater the similarity between the ontological types, the higher the score. Next, the relations describing the two are compared: having many relations in common yields a higher score but complete accordance is not obligatory. Finally, not only the number but also the kind of relations encoded is of importance: thus *petrol car* is more similar to *diesel car* than it is to *crane lorry* because although all three share the same relation HAS_PART, the relevant parts for the former two – *petrol engine* and *diesel engine*, respectively – belong to the same ontological type as opposed to the HAS_PART = ‘crane’ of the last.

Based on the similarity score, the co-hyponyms are presented in descending order. The list has been reduced to the 30 highest scoring words but with the possibility to see up to 200 as a clickable option.

The most problematic group is the list of hyponyms. As with the co-hyponyms, the list of hyponyms can be long but, unlike the co-hyponyms, we have found no meaningful automatic way of selecting the best candidates. If we look up the word *car*, should *petrol car* be considered more relevant than *crane lorry*? So far, we do not know the

answer and, as a provisional solution, we simply show a list of up to 200 words, randomly reduced. This is by no means expedient and for future updates we hope to develop a better method of presenting the information, for instance by grouping the hyponyms in relevant types.

2.1. Comparison with *Macmillan Online Dictionary*

A similar thesaurus function is offered by *Macmillan Online Dictionary* but in this case the contents have been manually edited. Figure 2 shows the thesaurus entry for 'car'.

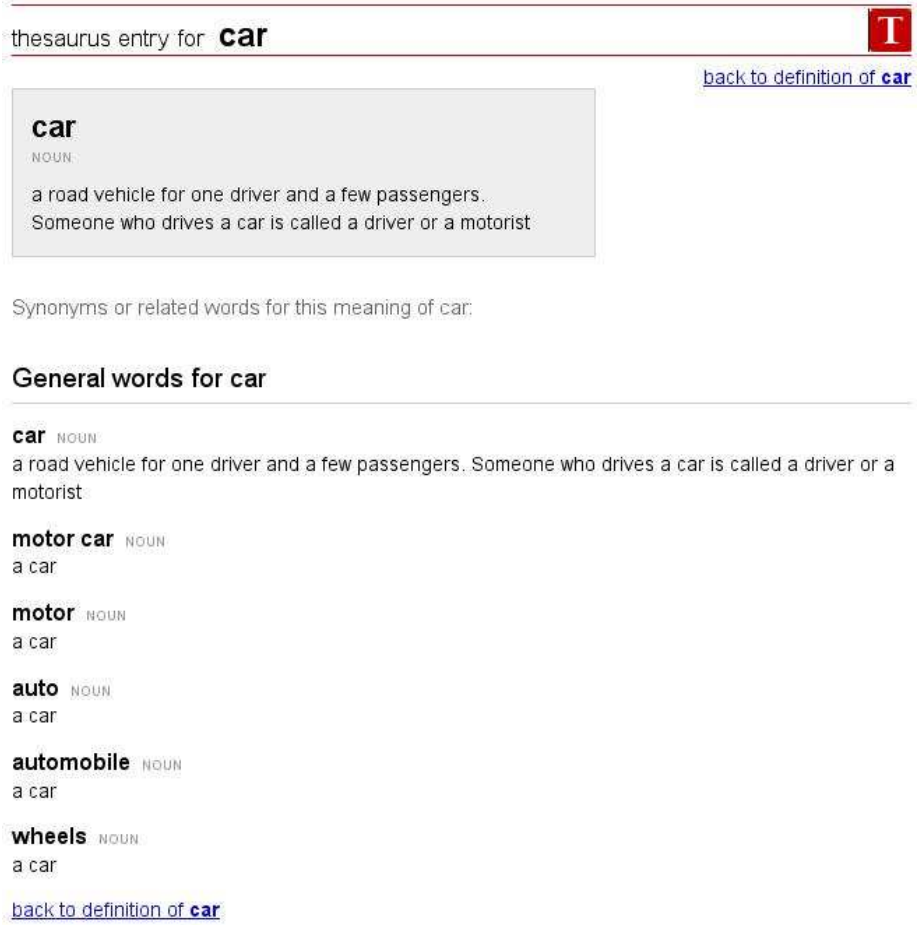


Figure 2. Macmillan's thesaurus entry for 'car'

This is an alternative way of doing things and it is instructive to compare the results. The first thing to notice is that the sheer number of words is much more manageable due to the fact that the words belong to more or less the same level of abstraction. In many cases this gives just the relevant alternatives for the user trying to find other

words in text production. Conversely, several of the examples in Figure 1, *e.g. overhead projector, oscillator* and *scanner*, are not likely ever to become real paradigmatic alternatives for *computer*. The problem of over-generation of candidates from DanNet is connected with the number of ontological types. Most thesauri from Roget onwards use 800-1,000 semantic groups whereas the norm of about 200 ontological types used in wordnets is bound to result in more members per group for a given vocabulary size – unless it is combined with other criteria, such as the ontological type of the target sense for a given relation. An example of an extreme case is the rich vocabulary connected with 'person'. Because of the lack of more subtle taxonomic subdivisions, a word like *catholic* has 3185 co-hyponyms, including *hippie, fascist, godfather, gourmet, ecologist* and *cat owner*, words that have little in common apart from the fact that they denote persons.

Another difference is that the heading of a superordinate term in DanNet as well as all the category members are always themselves words in the language. It is an essential feature of DanNet that it should reflect a "natural" categorization of the world, *i.e.* corresponding to the lexicalized labels for concepts of the Danish language. This is arguably the major difference between a wordnet and an ontology, and we deliberately wanted to avoid introducing conceptual categories that lack linguistic counterparts. We regard this as a strong point of DanNet, in particular with respect to its primary use in language technology. In thesauri for human users, however, the need for categories of a manageable size is more important than having categories with only single word headings. This remains a problem when it comes to the presentation of WorldNet data in comparison with the manually compiled thesaurus.

On the other hand, the manual approach has its problems, too. However commendable the effort to arrive at numerically manageable categories, it all depends on the meaningfulness of the headings chosen. Take *chinchilla* as a case in point: if you are looking for alternatives in Macmillan you arrive at the category "Mammals found in North, Central and South America", with 27 members. My guess is that you are as likely to be interested in other rodents or in other pets as you are in *alpaca, coyote, grizzly* or *caribou* as alternative words for *chinchilla*. Likewise, if you are looking for other words for 'off-road vehicle' under the heading "Vehicles used away from roads and on snow" you will not find *Land Rover* because this has been assigned to "Makes of car"; and *four-wheel drive* is found under "Equipment and systems in cars and other road vehicles", whereas *jeep* has been placed under "Military and industrial vehicles".

And although Macmillan is generally economical and to the point, they sometimes also face the problem of having too many or ill-suited words. If you look up the word *confectionery*, for example, you arrive at a group labelled "Types of food or drink", a sizeable group with almost 50 members but not many of them obvious alternatives for *confectionery*: *aphrodisiac, baby food, creole, macrobiotic, nutraceutical, slop* and *wholefood*, to name but a few. I hasten to emphasize that this is not a general impression: If, instead of *confectionery*, you look for *sweets*, the group "Sweets and

other confectionery" contains an equivalent number of, pardon the pun, very palatable examples.

In our case, the solution to the over-generation of hyponyms seems to be either to introduce sub-categories manually, especially towards the more abstract end of the semantic cline, corresponding to Macmillan's groups of "general words for *person*, *vehicle*, *machines*, etc.", or to develop a quantitative method that would allow us to rank and select the most appropriate words as it is done for co-hyponyms. But at the time of writing we have not accomplished this, which is why a small "beta" sign has been attached to the function label for "Related words".

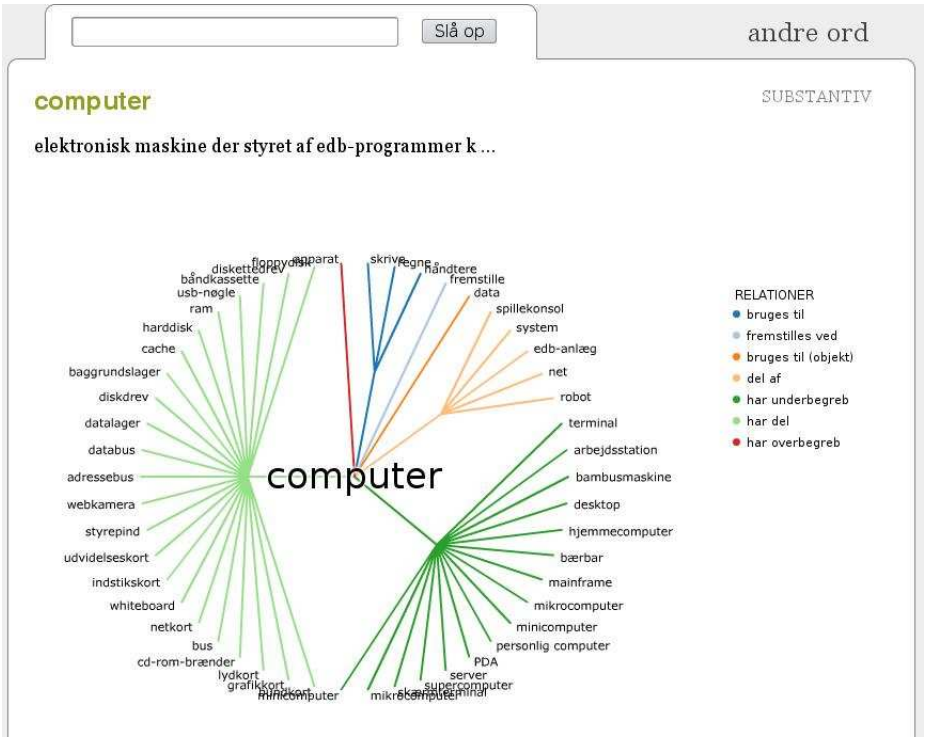


Figure 3. Visual representation of "computer" in andreord.dk

Finally, to meet the needs of language learners and others interested in systematic vocabulary training and semantic fields we have long wanted to bring an overall visual presentation of related words, along the lines of *The Visual Thesaurus* and other viewers that allow the user to browse wordnet data. But as sometimes happens when you release data as open source, others do the job for you. This is what happened for us when we discovered andreord.dk ('other words'), a site that does precisely that. So, instead we have chosen to link to this external resource for the visual presentation. An example is shown in Figure 3.

Apart from the visual presentation, the site also provides a search box, an extract from the definition, synonyms, one or more example sentences, the path of hyperonyms, hyponyms and the ontological type.

3. Dictionary and corpus data

Dictionary data are linked with corpus data in various ways. Figure 4 shows the entry for 'flag' (same word in English) in DDO. In the left column the user can:

- look up the relevant lemma in the corpus (1);
- calculate collocates of the lemma (2);
- look up the corresponding lemma in the ODS (3);

And, in the centre column, the user can

- look up corpus examples of the specific collocations shown (4).

For all three options in the left column, the user can choose whether the query should be for a) the string, or b) a particular part of speech. The latter is particularly useful in case of homonymy.

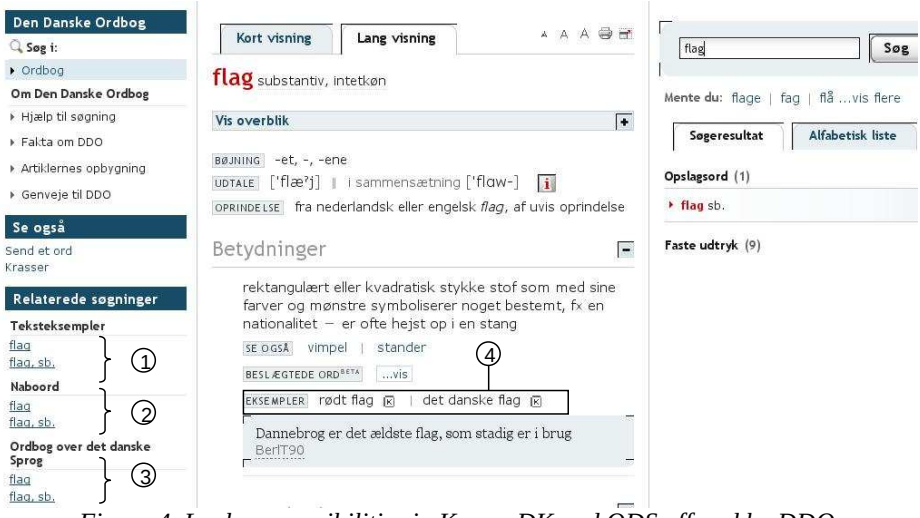


Figure 4. Look-up possibilities in KorpusDK and ODS offered by DDO

The "K" icon (K for *korpus*) in the centre column is clickable and when activated it submits a query for that collocation. The result for "rødt flag" (red flag) is shown in Figure 5 as a common concordance display with the searched words highlighted.[3]

[3] The default setting for multi-word queries allows up to three intervening words – hence the varying number of highlighted words – but the setting for multi-word queries can be customized at the user's will.

The alternative query options are, as always, found in the left column (under the heading “Relaterede søgninger” in Figure 5). Here it is possible to:

- calculate collocates of any of the constituting words, based on the possible lemma forms (1);
- get a list of fixed phrases containing any of or all the constituting words (2);
- look up the multi-word expression or any of the constituting words in the DDO, as a string or as a specific POS (3);
- look up any of the constituting words in the ODS, as a string or as a specific POS (4).

The list of fixed expressions in (2) is itself derived from DDO. It represents a subset of multi word expressions (bearing in mind that the DDO was itself based on evidence from a subset of KorpusDK), *viz.* those expressions that were selected for lemmatization by the editors during the manual compilation process. Concordances for words or expressions are easily generated with a click, both from the list of collocates and from the list of fixed expressions.

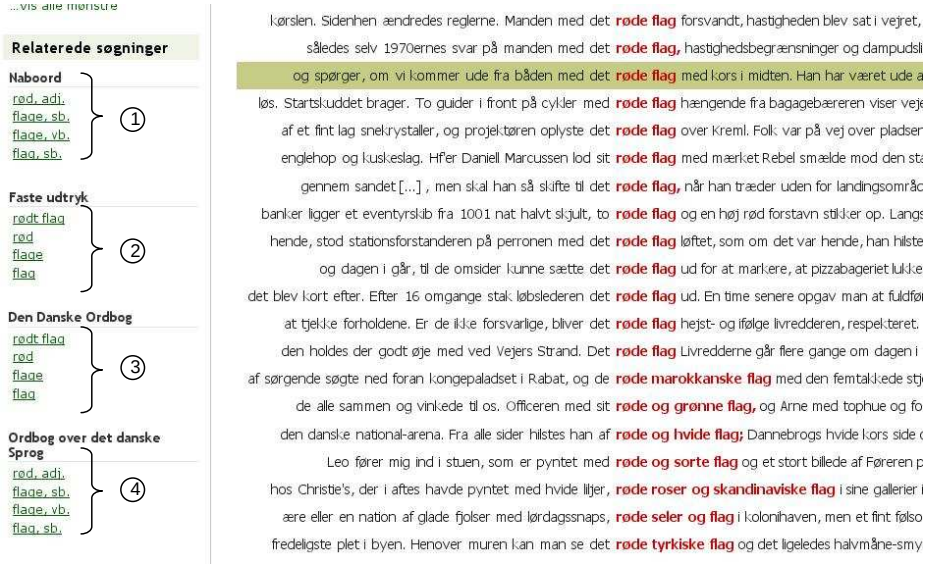


Figure 5. Look-up possibilities in KorpusDK, DDO and ODS

5. Perspectives

This is roughly the current state of affairs but, obviously, things do not stop here. Among the priorities for continued development are the following:

First, we need improved tools for tagging and parsing new corpus material. At present, the KorpusDK contains c. 56 million tokens and has not changed since it was released

in 2002. Texts have been collected continuously since 2005 but await mark-up. Provided we can develop new and expedient mark-up procedures, we hope to supply lemmas, variants and inflectional forms with corpus frequency information. With proper syntactic mark-up we would also be able to offer look-up possibilities for the valency patterns given in the grammatical section. And with a continuous influx of texts, corpus analysis can be automated to generate candidates for new lemmas in the dictionary.

Another perspective is the development of an integrated separate onomasiological presentation of the DanNet data where the user can query and navigate the semantic hierarchy, *e.g.* through a tree-structure view of the nearest superordinate and subordinate levels. Whether it should be via a separate search page or integrated in the entry (*i.e.* as the current presentation of "Related words") remains an open question at this point. Recently, The Society for Danish Language and Literature has received funding for a three year project to develop a traditional thesaurus based on data from DanNet. This will allow us to address the problems and shortcomings that have been pointed out here.

Finally, we would like to incorporate more grammatical information in the dictionary. Our institution is involved in the edition of a new comprehensive grammar of Danish and an obvious perspective is to link directly from the grammar sections of a dictionary entry to the relevant paragraph in the grammar.

References

ANDERSEN, M.S., ASMUSSEN, H. and ASMUSSEN, J. (2000). The Project of Korpus 2000 Going Public. In A. Braasch C. and Povlsen (eds). *Proceedings of the 10th EURALEX International Congress*. Volume 1, Copenhagen: Euralex: 291-299.

LORENTZEN, H. and TRAP-JENSEN, L. (2008). The Dictionary of the Danish Language Online: From Book to Screen – and Beyond. In *Lexicographie et Informatique – bilan et perspectives. Pré-actes*, ATILF / CNRS, Nancy-Université: 151-157.

Macmillan Online Dictionary. http://www.macmillandictionary.com

Ordnet.dk. http://ordnet.dk

PEDERSEN, B.S., NIMB, S., ASMUSSEN, J., SØRENSEN, N.H., TRAP-JENSEN, L. and LORENTZEN, H. (2009). DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. In *Language Resources and Evaluation*, vol. 43, no 3. Springer Netherlands: 269-299.

ROGET, P.M. (1998). *Roget's Thesaurus of English Words and Phrases*. 1998 edition by Betty Kirkpatrick. London: Penguin Books.

SØRENSEN, N.H. and TRAP-JENSEN, L. (forthcoming). Den Danske Ordbog som begrebsordbog. In H. Lönnroth and K. Nikula (eds). *Nordiske Studier i Leksikografi*. No. 10, NFL-skrift nr. 11, Tammerfors.

Showing phraseology in context: onomasiological access to lexico-grammatical patterns in corpora of French scientific writings

Agnès Tutin[1]
LIDILEM, Université Grenoble 3 – Stendhal

Abstract

Phraseology is prevalent in scientific writing and a good command of prefabricated patterns in this genre is required to write high quality texts and to be fully integrated into the "academic tribe". However, for French, few phraseological resources are currently available. For a writing aid in a second language, the user needs authentic corpus-based examples and an onomasiological or rhetorical access to this lexicon. NLP techniques offer an interesting perspective for this kind of application and a model of this kind; using rhetorical functions and semantic and syntactic properties, inspired by Sandor's concept matching (Sandor 2007) and the notion of frame (Fillmore *et al.* 2003), is presented here. A first implementation has been realized in the context of the Scientext project.

Keywords: phraseology, academic and scientific writing, collocations, NLP, corpora.

1. Introduction

As outlined by Gledhill (2000), Pecman (2004, 2007) and Loffler-Laurian (1987), phraseology is particularly prevalent in scientific writings, as in many professional discourses (*e.g.* marketing discourse, Cusin Berche 1998). There are both a specific lexicon and a specific phraseology of scientific writings, which are to a certain extent cross-disciplinary, and have been built especially for teaching applications in a second language. For example, Phal's Vocabulaire Général d'Orientation Scientifique (Phal 1971) is a basic scientific lexicon of scientific books within the framework of the *français fondamental* (Gougenheim *et al.* 1964), while in English Coxhead's Academic Word List (Coxhead 2000; Hirsh and Coxhead 2007) is intended to be a lexical database for teaching applications in English for Academic Purposes. More recently, several studies have used NLP techniques to define this cross-disciplinary scientific lexicon (Drouin 2007; Paquot and Bestgen 2009) and have proposed semantic descriptions of this lexicon and phraseology (Pecman 2004; Tutin 2007).

A good command of this lexicon is required, not only to write high quality texts (in order to be published and to get positive evaluations from peers), but also to use the

[1] agnes.tutin@u-grenoble3.fr

proper language of the "academic tribe" and to be fully integrated in the academic community, as outlined by Swales (1990). Phraseology is here defined in a broad sense as recurrent multiword expressions which are lexicalized. It is particularly important in academic language since it is used as a set of prefabricated "ready to use" patterns for the scientific writer. It is essential to provide a semantic and a rhetorical description of this phraseology, and to make available a large set of corpus-based examples of this specific lexicon. Resources of this kind are almost non-existent in French for academic language. However, NLP techniques when combined with semantic treatments, offer an interesting perspective for this kind of application.

In this paper, I will first present examples of corpus-based phraseology and several semantic treatments of this lexicon. I will then develop a proposal for an onomasiological and rhetorical access to this phraseology, and a first implementation within the framework of the Scientext project.

2. Corpus-based lexical resources and onomasiological approaches to phraseology

2.1. Corpus-based lexical resources

What non native speakers often need most as a writing aid is probably a large set of authentic corpus-generated examples, which illustrate the phraseology in its natural environment. They often have a passive knowledge of this vocabulary and just need to select the right collocations and check the syntactic constructions, which can vary considerably. Several tools showing collocations in context are already available with the help of NLP techniques. A well-known example is Kilgarriff's sketch engine (Kilgarriff *et al*. 2004), based on NLP techniques, which provides a corpus-based profile of a word (a "word sketch"). For example, for the word *hypothesis*, the most relevant collocates are sorted by syntactic structure. One can thus look up the most salient collocations in which *hypothesis* is the direct object, the subject, has a modifier, and other relations as shown on Figure 1. For the word *hypothesis*, the most significant verb-noun collocations are *to test a hypothesis* or *to support a hypothesis.*

It is also possible to display all the concordances of this collocation, including syntactic alternations such as passive voice or relative constructions.

A second very interesting corpus-based dictionary is the "dictionnaire des co-occurrences" of the French writing aid *Antidote* (Druide Company). This tool incorporates several dictionaries, including a dictionary of collocations. This lexical database is semi-automatically generated with the help of NLP techniques (Charest *et al.* 2007) and provides a list of collocations organized according to syntactic patterns and sorted by association measure. Relevant corpus examples, which have been manually checked, are associated to each collocation, as shown in Figure 2.

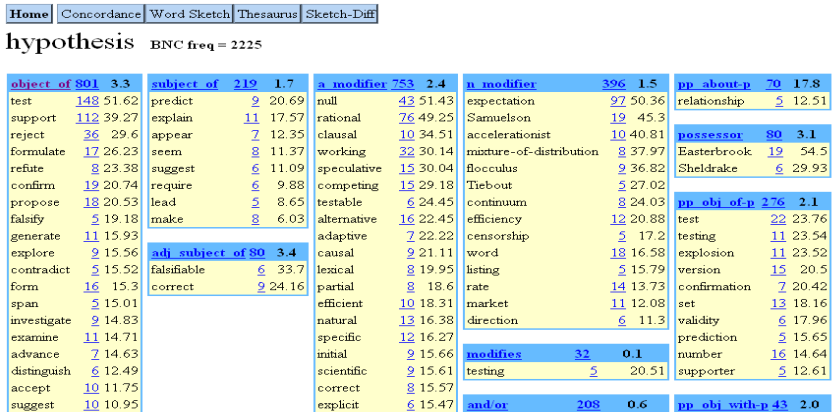


Figure 1. A Word Sketch of hypothesis in the Sketch Engine

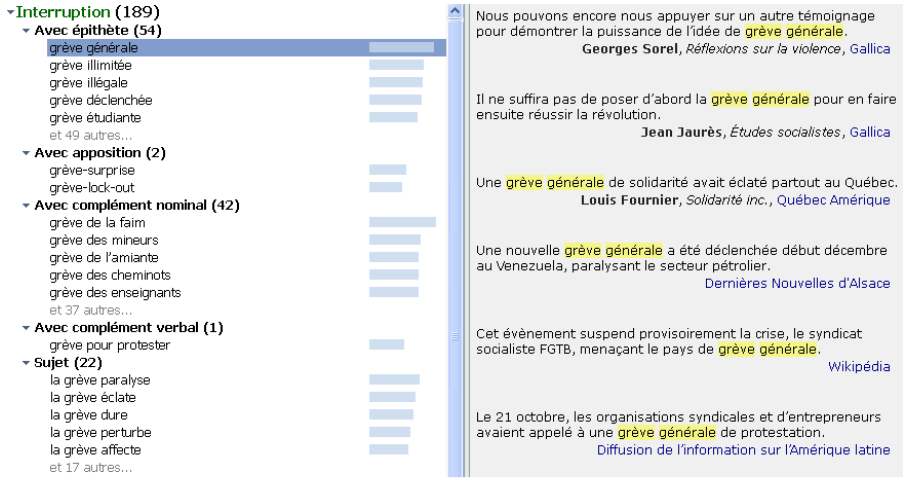


Figure 2. An example of Antidote' dictionary of collocations

For example, for the word *grève* (strike), we can find verb-noun collocations like *déclencher une grève* (to launch a strike) and authentic examples drawn from newspaper corpora. Nevertheless, in these two resources, there is regrettably no semantic description of the collocations, which is a problem for a writing aid tool in a foreign language.

2.2. Onomasiological approaches to phraseology

This section presents several semantic or rhetorical classifications used to describe the scientific phraseology. Such approaches, combined to corpus-based examples, seem very promising.

2.2.1. Pecman's ontology of the General Scientific Language

A first kind of typology has been designed by Mojca Pecman's "General Scientific Language" (Pecman 2004; 2007), a very complete and sophisticated ontology to account for the phraseology in this sublanguage. This ontology is organized under 4 main spheres (*i.e.* the scientific, universal, modality and discourse spheres) and includes 125 concepts. Each concept is related to a set of multiword units in French (and their translation in English). For example, for the concept 'collaboration', as shown in Figure 3, we have expressions such as *to encourage a collaboration, to establish collaboration...*

-/COLLABORATION/

collaboration *(phr.)* **apporter sa collaboration à/sur** : to collaborate on/in, ex. apporter sa collaboration à un projet : to collaborate [on/in] a project ; **développer davantage une collaboration** : to expand a collaboration ; **encourager une collaboration** : to [encourage/foster] a collaboration ; **établir**, **développer**, **mettre en place une collaboration** : to [establish/develop/build up] a collaboration ; **étudier <qch> en collaboration avec <qn>** : to study <sth> in collaboration with <sb>

Figure 3. Phraseological units associated to the concept of 'collaboration'

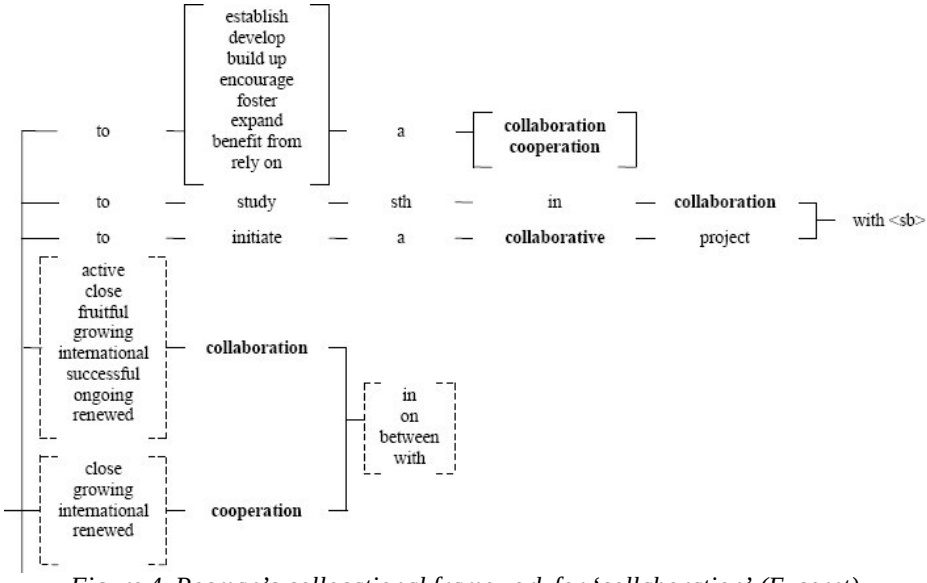


Figure 4. Pecman's collocational framework for 'collaboration' (Excerpt)

Pecman also proposes to represent multiword units in collocational frameworks (as already formalized by Altenberg 1998). According to the author, this representation makes it possible to account for both the syntagmatic and the paradigmatic levels. In the example for the concept of 'collaboration' (see Figure 4), we can see all the expressions and the variations related to this concept (for example, morphologic variations or lexical variations). This representation is very close to finite state automata and can be easily implemented in local grammars applied to corpora. Local grammars using dependency parsers are nevetheless more effective, because these syntactic representations are closer to semantic representations than surface automata.

2.2.2. Rhetorical approaches

The phraseology of scientific writing has also been analyzed within the framework of studies dealing with the concept of metadiscourse, whose classical definition is, according to Crismore and his colleagues:

> Linguistic material in texts, written or spoken, which does not add anything to the propositional content but that is intended to help the listener or reader organize, interpret, and evaluate the information given. (Crismore *et al.* 1993: 40)

These approaches are generally related to the rhetorical functions of phraseology in scientific writings where formulaic sequences are associated with specific goals, such as, for scientific writings:

- Giving evidence to a fact: *as we can see on this table, we can observe that*
- Structuring the text with metatextual markers: *We will first present …*
- Giving an opinion: *According to us … We think that …*
- Contrast and comparison with peer work: *Contrary to Parker (1990), we hypothesize … Our study differs from …*
- Scientific filiation or academic filiation: *Following Parker (1979) we hypothesized that, we borrow Parker's definition …*

Several models have been proposed to account for these rhetorical functions. A very popular model is Teufel's (1998) "argumentative zoning", originally used for automatic summarization. Teufel associates a list of formulaic expressions with specific rhetorical functions. For example, *to our knowledge* is associated with the function "gap introduction" while *when compared to* is associated with the function "comparison".

However, as highlighted by Sandor (2007), the formulaic expressions used by Teufel are just "bags of words" and are not organized from a syntactic and a semantic viewpoint. To fill this gap, Sandor proposes an alternative model, the "concept matching" model which includes a semantic and a syntactic representation for rhetorical functions. For example, the background discourse function is composed of concepts such as 'background' and 'knowledge' which in turn are decomposed into subconcepts such as 'general', 'past', 'researcher', 'mental activity' ... The concepts are related to a list of lexical items, and the model includes syntactic relations.

Sandor's perspective seems very effective and has been a source of inspiration for our project.

3. A proposal and an implementation for an onomasiological approach to extract corpus-based phraseology

3.1. Representation

The present proposal aims at providing the phraseology associated with specific rhetorical functions. The representation includes three levels:

- A **rhetorical level** indicates the rhetorical function, for example giving an opinion, or showing scientific affiliation.
- The **semantic and enunciative level** includes semantic predicates, the semantic roles of the participants, participants in the enunciative situation. Our representation here is close to Fillmore's semantic frames, or Sandor's "concept matching".
- A third level is the level of the **lexical markers** and of the **syntactic organization**.

I will illustrate this representation with the example of the rhetorical function of scientific affiliation, which is lexicalized into examples as the following ones:

1. Cette étude reprend ici les travaux de Bunt (1995) sur …
   [This study uses here the work of Bunt (1995) …]
2. A la suite de (Lee 1986), nous considèrerons que l'employeur … [Following (Lee 1986), we consider that …]
3. nous avons choisi d'utiliser le modèle des graphes conceptuels [Sowa, 1984] … [We chose to use the model of conceptual graphs (Sowa, 1984) …]

In spite of their lexical diversity, all these examples can be represented with the help of the same semantic frame, including four main concepts:

- The **agent**, generally the **author**, for example *we* or *I*, but also in a metonymic extension, *cette étude* ('this study').
- The **act of borrowing** a concept or a scientific object : *utiliser* ('use'), or *à la suite de* ('following').
- The **scientific concept** which is borrowed: *travaux* ('work') or *modèle* ('model').
- The **agent**, who is the **scientific source**, generally a proper noun followed by a date.

Figure 5 summarizes this semantic frame. From the syntactic and lexical viewpoint, several constructions can be associated to this frame, as represented in Figure 6, with the help of a dependency-like representation.

In sentences like *à la suite de X* ... (‘following X’), the concept of ‘borrowing’ is a preposition and the source is a prepositional complement, while in constructions like *nous utilisons le modèle de X* (‘we use X’s model), the author is the subject and the concept of ‘borrowing’ is a verb. It is possible to list the various syntactic structures where this frame can be encountered.

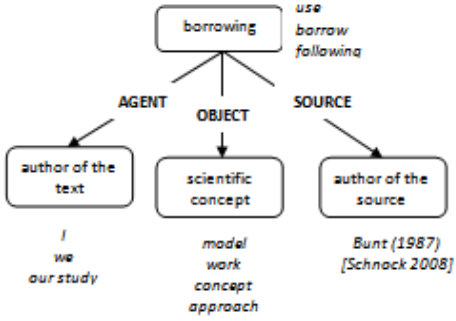


Figure 5. A semantic frame for the rhetorical function of scientific affiliation

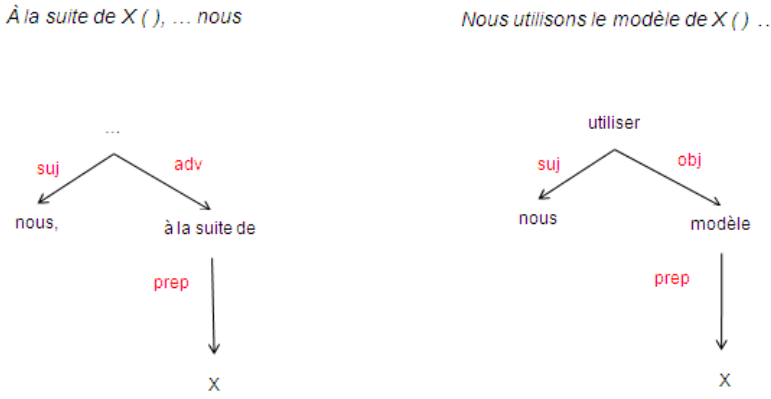


Figure 6. Several syntactic configurations for the scientific affiliation rhetorical function

Several local grammars have been implemented following this principle, where several syntactic configurations are associated with rhetorical functions, including a set of quasi-synonyms. Some syntactic dependency relations are computed into semantic relations. For example, active, passive, and pseudo-passive constructions can be merged into a deep-object relation, as proposed by Hagège and Roux (2003).

3.2. A first implementation within the Scientext Project

A first implementation of these local grammars has been realized in the context of the Scientext project. This project, funded by the French ANR involves several teams (Grenoble's LIDILEM[2], Chambéry's LLS[3] and Lorient's LiCorn[4]) and aims at building a freely-available corpus of academic writing in French and English, and devising tools for studying linguistic markers and phraseology of stance/positioning and reasoning.[5]

Specific tools have been designed to search the corpus. The interface has been created by Achille Falaise, from Olivier Kraif's Concquest query language (Kraif 2008). The corpus has been syntactically analyzed with the dependency parser Syntex (Bourigault 2007).

The user can first select the corpus according to several parameters, such as the discipline (or the family of disciplines), the textual genre (research article or thesis), and the textual part (the corpus is TEI-Lite annotated), as presented in Figure 7.

Figure 7. Corpus selection in Scientext

Once the corpus is selected, the user can access the data via several interfaces. The simple interface provides scroll menus with forms, lemmas and syntactic categories, and makes it possible to use the syntactic dependencies in a straightforward manner. For example, on Figure 8, the query will extract verbs having the lemma *hypothèse* as a direct objet.

[2] Laboratoire de Linguistique et de Didactique des Langues Etrangères et Maternelles (www.u-grenoble3.fr/lidilem).
[3] Laboratoire Langages, Littératures, Sociétés (http://www.lls.univ-savoie.fr/).
[4] Linguistique de Corpus (équipe LiCorn : http://web.univ-ubs.fr/corpus/index.html).
[5] www.u-grenoble3.fr/lidilem/scientext

Figure 8. A query with the simple interface

A complex query language can be used, using part of speech tagging and syntactic dependencies, combined with variables in a more sophisticated way than the simple queries presented above. Predefined local grammars dealing with stance and positioning, built according to the model presented in the previous section, make it possible to search for the phraseology associated with a rhetorical function. These grammars include variables and semantic relations (for example, passive and active constructions, *e.g. this confirmed the hypothesis* or *the hypothesis was confirmed* are considered the same semantic construction,). Lexical sets of quasi-synonyms are associated with concepts. Figure 9 shows a simplified grammar for scientific affiliation, where the concepts 'author', 'scientific object' and 'borrowing' are used.

```
// Grammaire de la filiation

// Objet emprunté

$objet=(<lemma=analyse,#3>|<lemma=approche,#3>|<lemma=concept,#3>|<lemma=définition,#3>|<lem
ma=démarche,#3>|<lemma=démonstration,#3>|<lemma=étude,#3>|<lemma=hypothèse,#3>|<lemma=m
esure,#3>|<lemma=méthode,#3>|<lemma=modèle,#3>|<lemma=notion,#3>|<lemma=procédure,#3>|<le
mma=statistique,#3>|<lemma=terme,#3>|<lemma=terminologie,#3>|<lemma=théorème,#3>|<lemma=thé
orie,#3>|<lemma=travail,#3>) ;

// Auteur

$auteur =(<form=nous,#1>|<form=on,#1>|<lemma=je,#1>) ;

// Verbe d'emprunt

$vemprunt =
(<lemma=choisir,#2>|<lemma=compléter,#2>|<lemma=employer,#2>|<lemma=étendre,#2>|<lemma=exp
loiter,#2>|<lemma=généraliser,#2>|<lemma=poursuivre,#2>|<lemma=reprendre,#2>|<lemma=retenir,#2>
|<lemma=suivre,#2>|<lemma=utiliser,#2>);

Main = $auteur && $vemprunt  && $objet :: (SUJ,#2,#1)(OBJ,#2,#3) ;
```

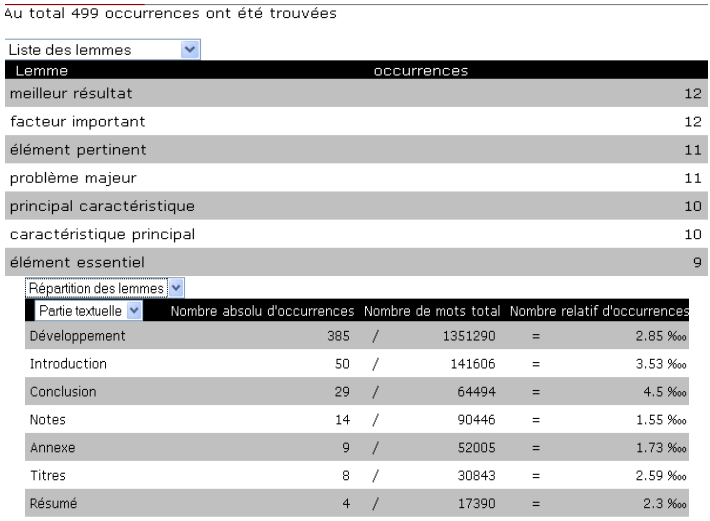
Figure 9. A simple local grammar of scientific affiliation

Results of the query are displayed in KWIC concordances (or larger text units) and some simple statistics have been developed as can be seen on Figure 10. For example, with the help of a predefined grammar of evaluative associations, the most frequent evaluative adjectives are displayed in every discipline, textual genre or textual part. For example, we have shown that evaluative adjectives are more numerous in

introductions and conclusions where persuasive strategies are more visible than in more technical sections (Tutin, to appear).

4. Conclusion

For an onomasiological access to phraseology in a second language in a writing aid software, NLP techniques offer interesting perspectives. They make it possible to build local grammars, implementing recurrent expressions which can be regularly associated with semantic and rhetorical functions. Like Sandor (2007), I think that the description of this phraseology should be organized according to semantic, syntactic and lexical principles, in order to avoid *ad hoc* descriptions. A first implementation has been realized in this direction in the context of the Scientext project. However, much work remains to be done to offer a complete writing aid. The linguistic model also still needs to be improved.


Au total 499 occurrences ont été trouvées

Liste des lemmes

Lemme	occurrences
meilleur résultat	12
facteur important	12
élément pertinent	11
problème majeur	11
principal caractéristique	10
caractéristique principal	10
élément essentiel	9

Répartition des lemmes

Partie textuelle	Nombre absolu d'occurrences		Nombre de mots total		Nombre relatif d'occurrences
Développement	385	/	1351290	=	2.85 ‰
Introduction	50	/	141606	=	3.53 ‰
Conclusion	29	/	64494	=	4.5 ‰
Notes	14	/	90446	=	1.55 ‰
Annexe	9	/	52005	=	1.73 ‰
Titres	8	/	30843	=	2.59 ‰
Résumé	4	/	17390	=	2.3 ‰

Figure 10. Simple statistics for the query on evaluative adjectives[6]

References

ALTENBERG, B. (1998). On the phraseology of Spoken English: The Evidence of Recurrent Word-Combinations. In A.P. Cowie (ed.). *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press: 101-122.

[6] If queries include lemmas, they are displayed on the interface.

BOURIGAULT D. (2007). *Un analyseur syntaxique opérationnel : SYNTEX*. Mémoire d'Habilitation à Diriger les Recherches, Toulouse.

CHAREST S., BRUNELLE E., FONTAINE J. and PELLETIER B. (2007). Élaboration automatique d'un dictionnaire de cooccurrences grand public. In F. Benmara, N. Hathout, Ph. Muller and S. Ozdowska (eds). In *Actes de TALN 2007*, Toulouse: 282-292.

COXHEAD, A. (2000). A New Academic Word List. *TESOL Quarterly*, 34 (2): 213-238.

CRISMORE, A., MARKKANEN, R. and STEFFENSEN, M. (1993). Metadiscourse in persusasive writing: A study of texts written by American and Finnish university students. *Written Communication*, 10(1): 39-71.

CUSIN-BERCHE, F. (1998). L*e Management par les mots: étude sociolinguistique de néologie*. Paris: L'Harmattan.

DROUIN, P. (2007). Identification automatique du lexique scientifique transdisciplinaire. *Revue française de linguistique appliquée*, 12(2): 45-64.

FILLMORE, Ch., JOHNSON Ch. and PETRUCK M. (2003) Background to Framenet? *International Journal of Lexicography*, 16(3): 235-250.

GLEDHILL, Chr. (2000). *Collocations in Science Writing*. Tuebingen: Gunter Narr Verlag (*Language in performance*, 22).

GOUGENHEIM, G., MICHA, R., RIVENC, P and SAUVAGEOT, A. (1964). *L'élaboration du français fondamental (1º degré)*. Paris: Didier.

HAGÈGE C. and ROUX C. (2003). Entre syntaxe et sémantique: Normalisation de la sortie de l'analyse syntaxique en vue de l'amélioration de l'extraction d'information à partir de textes. *Actes de TALN, TALN 2003*, Batz-sur-Mer, 11-14 juin 2003.

HIRSH, D. and COXHEAD, A. (2007). A pilot science-specific word list. *Revue Française de Linguistique Appliquée*, XII(2): 65-78.

KRAIF O. (2008). Comment allier la puissance du TAL et la simplicité d'utilisation ? l'exemple du concordancier bilingue ConcQuest, In *Actes des 9e Journées d'analyse statistique des données textuelles, JADT 2008*. Lyon: Presses universitaires de Lyon: 625-634.

KILGARRIFF A, RYCHLY, P., SMRZ, P. and TUGWELL, D. (2004). The Sketch Engine. In G. Williams and S. Vessier (eds). *Proceedings of Euralex*. Lorient: 105-116.

LOFFLER-LAURIAN, A.-M. (1995). Locutions et discours scientifiques. In M. Martins-Baltar (ed.). *Cahiers du Français Contemporain*, La locution en discours, 2, Décembre 1995: 243-269.

PAQUOT, M. and BESTGEN, Y. (2009). Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. In M. Hundt, D. Schreier and A.H. Jucker (eds). *Corpora: Pragmatics and Discourse*. Amsterdam: Rodopi: 243-265.

PHAL, A. (1971). *Vocabulaire général d'orientation scientifique (V.G.O.S.) – Part du lexique commun dans l'expression scientifique*. Paris: Didier, Crédif.

PECMAN, M. (2004). *Phraséologie contrastive anglais-français : analyse et traitement en vue de l'aide à la rédaction scientifique*. Thèse de doctorat, Université de Nice Sophia Antipolis, décembre 2004.

PECMAN, M. (2007). Approche onomasiologique de la langue scientifique générale. *Revue Française de Linguistique Appliquée*, XII(2): 79-96.

SÁNDOR, A. (2007). Modeling metadiscourse conveying the author's rhetorical strategy in biomedical research abstracts. *Revue Française de Linguistique Appliquée*, XII(2): 97-108.

SWALES, J.M. (1990). *Genre Analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

TUTIN, A. (2007). Modélisation linguistique et annotation des collocations : une application au lexique transdiciplinaire des écrits scientifiques. In S. Koeva, D. Maurel and M. Silberztein (eds). *Formaliser les langues avec l'ordinateur : de Intex à* NooJ. Besançon: Presses Universitaires de Franche Comté: 189-215.

TUTIN, A. (to appear). Evaluative adjectives in academic writing in the humanities and social sciences, *Interpersonality in written academic discourse: perspectives across languages and cultures*. Cambridge: Cambridge Publishing.

Software

Antidote HD. Druide Company. Montréal. http://www.druide.com.

SciE-Lex: an electronic lexical database for the Spanish medical community

Isabel Verdaguer[1], Elisabet Comelles, Natàlia J. Laso,
Eva Giménez, Danica Salazar
University of Barcelona

Abstract

The aim of this paper is to show the potential of SciE-Lex, a lexical database of non-specialised (bio) medical terms, intended to help Spanish scientists, especially those in the medical community, to write native-like scientific articles in English. SciE-Lex provides morphological, semantic, syntactic and collocational information. In line with new trends in corpus and phraseological studies, which emphasise the importance of controlling multi-word expressions as a device to structure discourse and improve language fluency, SciE-Lex is now being supplemented with 3- to 5-word clusters and provides explicit information about their composition, function and textual distribution. Formulaic constructions are classified according to their function in the discourse. The new additions in SciE-Lex also illustrate the advantages that the electronic format can offer, such as the possibility of connecting formulaic constructions to headwords by means of hyperlinks and the inclusion of information on the variability of prototypical expressions, accompanied by notes clarifying their usage patterns.

Keywords: EAP, health science, lexical database, discourse analysis, phraseology.

1. Introduction

Phraseological studies (*e.g.* Sinclair 1991, 2004; Howarth 1996, 1998; Granger 1998; Moon 1998; Hunston and Francis 2000; Wray 2002; Oakey 2002a, 2002b; Biber 2006; Hyland 2008; Meunier and Granger 2008) have provided strong evidence of the central role phrases play in language and shifted the focus of linguistic analysis from isolated words to multi-word units of meaning.

In addition, as English is now the 'lingua franca' in science and scholarship, mastery of the language is crucial to those who wish to disseminate their research to the international scientific community. The specialised vocabulary of the field does not pose many problems, since scientific and technical terms are very similar in different languages; however, non-natives may lack a good command of the general terms used in science and the prototypical phraseological conventions of the genre. This lack of

[1] i.verdaguer@ub.edu

phraseological knowledge, characteristic of non-natives, is also shared by junior native scientists (Gledhill 2000a, 2000b).

To provide autonomy to non-native scientists writing in English, English for Academic Purposes (EAP) courses are taught in an increasing number of universities, writing guides are being published (*e.g.* Swales and Feak 2005) and a section has been included in the second edition of the *Macmillan English Dictionary for Advanced Learners* (2007), aimed at improving the writing skills of learners in academic and professional contexts. However, a phraseological approach has not yet been used in specialised dictionaries.

Existing technical and scientific monolingual dictionaries usually lack contextual information on the syntactic and collocational patterns of non-technical vocabulary and on the usage of multiword units in English. It is thus necessary to develop lexical databases and specialised dictionaries that include lexico-grammatical and phraseological information, which can help to improve non-native scientists' performance in specialised English.

The aim of this paper is to show the potential of *SciE-Lex*, a lexical database of non-specialised (bio)medical terms, which is intended to help Spanish scientists, especially those in the medical community, to write native-like scientific articles in English. *SciE-Lex* provides explicit guidelines on the use of non-terminological lexical items and the word sequences they are typically found in.

The development of *SciE-Lex* in electronic format illustrates the advantages that this format can offer. It allows targeted search facilities that are not possible in print and improves access to required information. The electronic format makes it possible to connect formulaic constructions to headwords by means of hyperlinks and to include information on the variability of prototypical expressions. However, the selection of the content and the presentation of the information that the user would need in order to produce a native-like text is still a difficult issue, which we will approach in this paper.

2. Methodology

The information included in the dictionary has made it necessary to integrate corpus and discourse analysis. The main corpus analysed in this study is the *Health Science Corpus* (*HSC*), a restricted-domain corpus which contains a representative sample of texts specifically assembled for the current investigation of the use of non-specialised terms by the health science community. Given our interest in the lexico-grammatical patterns of non-technical terms in scientific English and the conventionalised phraseological characteristics of that genre, and the lack of publicly available scientific English corpora at that time, we decided to compile our own micro-corpus, which now consists of approximately 4 million words of scientific research articles from prestige online journals that cover different disciplines such as medicine, biology, biochemistry and biomedicine (for a complete description of the corpus see Laso 2009).

The results of our corpus analysis were codified in *SciE-Lex*, which, in its first stage, showed the combinatorial potential of words commonly used in scientific registers. *SciE-Lex* entries contain the following information: Word class, Morphological variants, Equivalent(s) in Spanish, Patterns of occurrence, List of collocates, Examples of real use and Notes to clarify usage. We illustrate the database with the word *evidence* (*cf.* Figure 1).

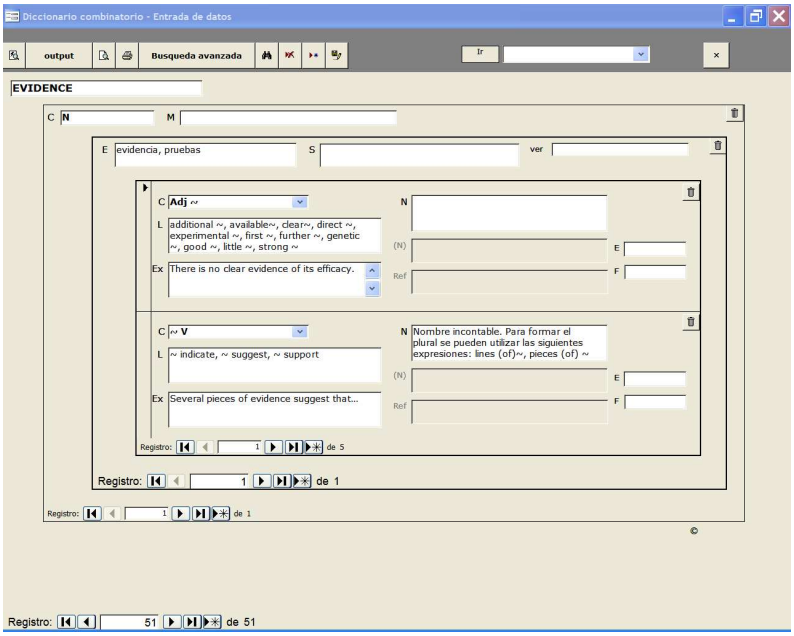


Figure 1. Lexical entry of the node word evidence

The software used in retrieving data from the *HSC* and refining the results further was version 5.0 of the concordancing programme *WordSmith Tools* (Scott 2008) and *ConcGram* (Greaves 2009). We used *Word Smith Tools* to make a list of general terms most frequently used in scientific English and to extract bundles automatically. We then made us of *ConcGram* to search for co-occurrences of two or more items irrespective of constituency and their positional variation. This enabled us to study the phraseological variability of the items under analysis.

In order to perform our studies on multiword units and select those to be included in *SciE-Lex* several steps were followed. First, we identified and selected possible candidates by analyzing different lengths of bundles with *WordSmith Tools*. For the purpose of this paper, we examined in particular detail the most common 3-, 4- and 5-grams in which the noun *evidence* occurs and concluded that:

- 5-word bundles consist of 4-word bundles and a variable element (*e.g. there is some/direct/further/good/increasing evidence that*)

- 3-word bundles tend to be included in 4-word bundles (*e.g. (provide) strong evidence that, (we) provide evidence that*), although others were unique (*e.g. body of evidence, lines of evidence*)

Following Biber *et al*. (1999), Cortes (2004) and Hyland (2008), we decided to include 5- and 4-word bundles and 3-word bundles which were not included in 4-word sequences. In this study, a minimum frequency of 5 occurrences per million words was set as the frequency cut-off point.

Secondly, we carefully examined the automatically generated list and removed those bundles that did not have meaning. Finally, we examined and analysed the selected bundles in terms of their variability, discourse function and distribution inside the text.

All this phraseological information is to be included in our *SciE-Lex* database, as it is considered to be of special relevance to non-native scientists, who must use lexical items in contextually appropriate ways in order to show phraseological competence in their scientific publications. This linguistic information should be clear and easy to access. Thus, we decided to make use of hyperlinks inside the entries. Bundles will be included in *SciE-Lex* entries, which will contain a hyperlink that will display all the information on variability, discourse function and distribution referring to each word sequence (*cf.* section 4).

3. Results and Discussion

The analysis of the *HSC* has revealed that the noun *evidence* participates in a wide range of recurring 4-and 5-word bundles. As already stated in the Methodology section, the use of corpus querying tools, such as *WordSmith 5.0* (Scott 2008) and *ConcGram 1.0* (Greaves 2009) has been extremely useful to uncover automatically the collocational and colligational environment of the word *evidence* in the *HSC*. This way of approaching the dataset helped us to conduct a qualitative analysis of concordance lines.

A careful examination of concordance lines shows that the following patterns can be formulated: *evidence* + verb + *that*-clause, adjective + *evidence* + prep. / *that*-clause and *there* + *be* + (neg. particle) + (adj./quantif.) + *evidence* + (prep. / *that*-clause). To illustrate the information provided by *SciE-Lex*, the pattern *there* + *be* + (neg. particle) + (adj./quantif.) + *evidence* + (prep. / *that*-clause) has been analysed with respect to its variability, textual distribution and functions.

As shown in Table 1, clusters may have different degrees of internal variation, a fact which has to be taken into account in the dictionary. The basic pattern *there* + *BE* + *evidence* can be followed by a preposition (*for*, *of*), an appositive *that*-clause or a *to*-infinitive clause. It can also be preceded by a negative particle, a quantifier (usually *some*) and/or an adjective.

Existential *there*	***be***	**Neg. particle**	**Adj./Quantif.**		**Prep./ *that*-Clause / *to*-inf.**	**Freq**
there	is					**43**
there	is				that	26
there	is				*to*-inf.	8
there	is		some	**evidence**		**14**
there	is		some		that	6
there	is		increasing			5
there	is		good			5
there	is		strong/little		that/of/for	5
there	is	no				**42**
there	is	no			that	17
there	is	no		**evidence**	for	15
there	is	no			of	5
there	is	no	direct/clear			7
	IS					**85**
there	was	no				**35**
there	was	no		**evidence**	of	24
there	was	no			that	7
	WAS					**35**

Table 1. Variability of the bundle there + BE + (neg. particle) + (adj./quantif.) + evidence + (prep./ that-clause) in the HSC

We found that instances of the noun *evidence* premodified by an adjective were also worth examining, as they seemed to be extended versions of the 4-word bundle *there* + *be* + *evidence* + prep. / *that*-clause.

In the analysis of the bundles, tense needs to be considered. The close observation of its variant forms reveals that the verb *to be* is only inflected either for the present form *is* (85 occurrences) or, to a lesser extent, in the past form *was* (35 occurrences). The interaction between tense and polarity is also worth noting. All occurrences in the past are used in the negative form (*i.e. there was no evidence of/that*), whereas the occurrences in the present tense are almost equally used both in the affirmative (43 occurrences) and negative forms (42 occurrences). The actual frequencies of this bundle are shown in Table 1.

Regarding the location of the bundle *there* + *be* + (neg. particle) + (adj./quantif.) + *evidence* + (prep. / *that*-clause) in the different sections and/or moves characteristic of the academic research article, findings have shown that this string is more commonly used in the Discussion section. Despite the fact that several occurrences are also traced both in the Introduction and Results section, Figure 2 below shows that the Discussion part embraces a wider range of variant forms of this bundle.

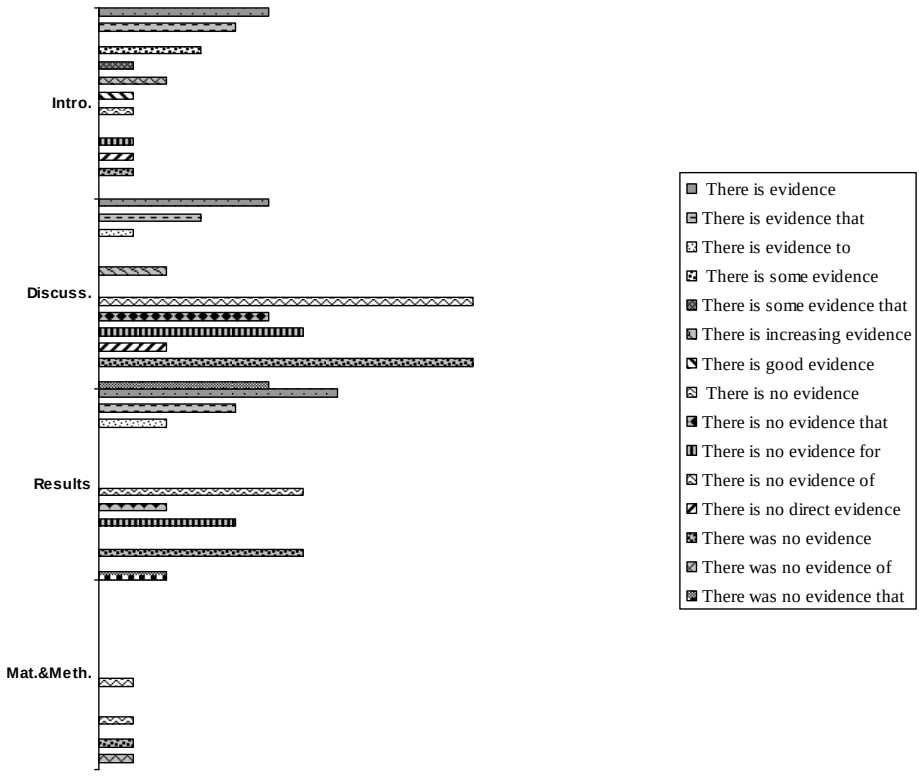


Figure 2. Variant forms and distribution across sections in the HSC.

In our analysis of the discourse function of the bundles, we have used Hyland's (2008) classification but we have also created some new labels intended to be more user-friendly and meaningful to our users, who do not necessarily have any background on linguistics.

With respect to the bundle *there* + *be* + (neg. particle) + (adj./quantif.) + *evidence* + (prep. / *that*-clause), three main functions can be identified: drawing conclusions, adding information and showing results. Nonetheless, we should highlight that the semantic load of the adjective plays an important role when defining the discursive function of the bundle. We can distinguish here evaluative adjectives (*e.g. strong*,

clear, good, conclusive, little), adjectives conveying the meaning of addition (*e.g. further*, *additional* and *increasing*) and other adjectives modifying the noun *evidence* such as *experimental* and *direct*.

Regarding evaluative adjectives, it is worth noting that adjectives which are semantically related may behave in different ways. Whereas clusters containing *strong* and *good* convey an assertive tone and fulfil the discourse function of drawing conclusions (*cf.* examples 1 and 2), the bundles containing the adjective *clear* are used with a negative form conveying tentativeness and fulfilling the function of showing results (*cf.* example 3).

1. For females, **there is strong evidence that** the accumulation of spontaneous mutations generates significantly greater genetic variation for mortality rates early in life than at older ages. (*HSC*).[2]
2. It is obvious that the moment in evolutionary history for each of the different organisms described here undoubtedly occurred not only as separate events but at different times. **There is good evidence that** multicellularity in cyanobacteria was invented at a very early stage, a good 3.5 billion years ago, 16 yet the first multicellular animals was probably a much more recent event (*HSC*).[3]
3. Cyclin D3 is a candidate for such a second target gene; while **there is no clear evidence of** functional distinction between cyclins D1 and D3, it is possible that they target cdk4, the catalytic subunit, to distinct substrates, and therefore might have non-redundant functions (*HSC*).[4]

Other bundles such as *there is increasing* / *further / additional evidence that/ for* are used to add more information to the previous data, which is clearly related to the quantitative meaning of the adjective, as can be seen in example 4.

4. **There is increasing evidence that** HSPGs can modulate the activity of growth factors through a number of mechanisms, including facilitating their dimerization or altering their effective concentration by acting as low-affinity receptors (*HSC*)[5].

4. Summary and conclusions

This paper highlights the importance of phraseological studies in dictionary making and shows the improvements which are achieved in our *SciE-Lex* database, by means of the introduction of 3 to 5-word clusters and the information associated with them, regarding their composition, function and distribution inside the text. In order to show the improvements which are being carried out in *SciE-Lex*, the pattern: *there* + *be* + (neg. particle) + (adj./quantif.) + *evidence* + (prep. / *that*-clause) has been explored.

Some interesting conclusions have been drawn from the analysis of these clusters. With regard to the discourse function of bundles, we agree with Oakey (2009) that bundles can be multifunctional, as we found some of them to have more than one

[2] *Genetics*, 153: 817.
[3] Integrative Biology, 27.
[4] *The EMBO* Journal, 18: 195.
[5] Genes and Development, 12, 12: 1901.

discourse function. Qualitative studies of the bundles are thus needed to include relevant information in the dictionary. In some cases, multifunctionality is related to the variability of the word sequence, since the function depends on the verb or the adjective involved, but in many cases the same cluster can have different functions.

Another aspect which needs to be mentioned is the relevance of hedging even in the presentation of evidence, thus corroborating Hyland's (2009) statement that hedges are frequently used in professional texts. Although it is true that assertiveness may be present, it is to be noticed that in some of the bundles presented here there is an adjective which indicates tentativeness. It is also worth noticing that the semantic load of the evaluative adjective used in bundles plays an important role when defining the discourse function of the bundle and this is clearly due to the assertive or tentative tone of the piece of text where they appear.

Polarity is another issue that has to be taken into account. It is relevant to highlight the strong relationship existing between the polarity of the bundles and the verb tenses included in these clusters, as can be seen in the bundle *there is/was no evidence that* in opposition to *there is evidence that*. Clusters with a positive polarity only contain verbs in the present tense, whereas negative polarity can be found with both present and past tenses. Through the use of negative polarity, two adjectives which could be considered near-synonyms, such as *clear* and *strong*, are used in different ways. Whereas *strong* is always used to indicate assertiveness, a high percentage of the occurrences of *clear* are linked to a negative element. Thus, although *clear* can certainly be used to indicate assertiveness, when it collocates with a negative element, it is used as a hedge.

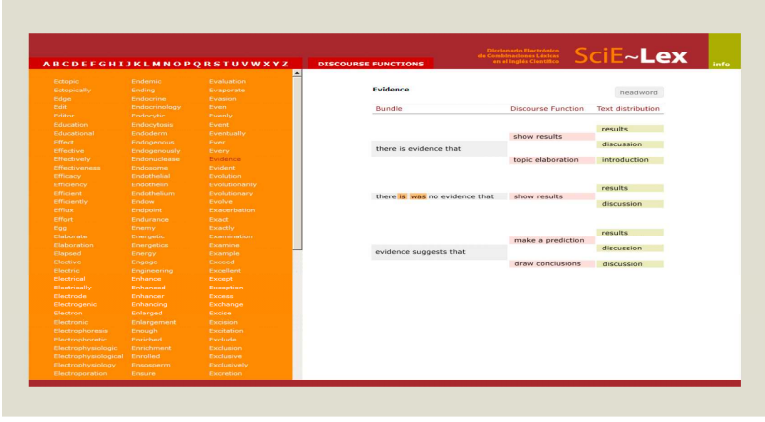


Figure 3. Output of the bundles containing the headword evidence in SciE-Lex

The information obtained from the analysis of the clusters regarding their variability, discourse function and textual distribution, together with other important aspects such as polarity, is considered to be of a special relevance to our dictionary users. This justifies the inclusion of this information in our *SciE-Lex* database by means of a "Bundle" button in the entry screen. By clicking on this button, the user is redirected to another page where the bundles appear on the left of the screen. Within the clusters, information including variability (*i.e.* different verb tenses or different verb forms), the possible discourse functions and text distribution is provided (*cf.* Figure 3). Other important information such as polarity will be included in a box below the bundle concerned.

By means of the introduction of bundles and their contextual information, we aim at providing non-native English speakers with enough tools and knowledge to gain control of the phraseological conventions of the medical genre in the English language.

Acknowledgments

The authors acknowledge the support of the *Ministerio de Educación y Ciencia* and FEDER (project number HUM2007-64332/FILO).

References

BIBER, D., FINEGAN, E., JOHANSSON, S., CONRAD, S. and LEECH G. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.

BIBER, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

CORTES, V. (2004). Bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, 23: 397-423.

GLEDHILL, C.J. (2000a). The discourse function of collocation in research article introductions. *English for Specific Purposes*, 19: 115-135.

GLEDHILL, C.J. (2000b). *Collocations in science writing*. Gunter Narr: Tübingen.

GRANGER, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A.P. Cowie (ed.). *Phraseology: Theory, analysis and applications*. Oxford: Oxford University Press: 145-160.

GREAVES, C. (2009). *ConcGram 1.0: A phraseological search engine*. Amsterdam: John Benjamins.

HOWARTH, P.A. (1996). *Phraseology in English academic writing: Some implications for language learning and dictionary making*. Tübingen: Max Niemeyer Verlag.

HOWARTH, P.A. (1998). The phraseology of learners' academic writing. In A.P. Cowie (ed.). *Phraseology: Theory, analysis and applications*. Oxford: Clarendon Press: 161-186.

HUNSTON, S. and FRANCIS, G. (2000). *Pattern Grammar*. Amsterdam: John Benjamins.

HYLAND, K. (2008). As can be seen: Bundles and disciplinary variation. *English for Specific Purposes*, 27: 4-21.

HYLAND, K. (2009). *Academic discourse*. London: Continuum.

LASO, N.J. (2009). A corpus-based study of the phraseological behaviour of abstract nouns in medical English: A needs analysis of a Spanish medical community. Unpublished PhD dissertation, University of Barcelona, Barcelona, Spain.

Macmillan English Dictionary for Advanced Learners. (2007). Oxford: Macmillan.

MEUNIER, F. and GRANGER, S. (eds) (2008). *Phraseology in foreign language learning and teaching*. Amsterdam and Philadelphia: John Benjamins.

MOON, R. (1998). *Fixed expressions and idioms in English: A corpus-based approach*. Oxford: Oxford University Press.

OAKEY, D. (2002a). A corpus-based study of the formal and functional variation of a lexical phrase in different academic disciplines. In R. Reppen *et al.* (eds). *Using corpora to explore linguistic variation*. Philadelphia: John Benjamins: 111-129.

OAKEY, D. (2002b). Lexical phrases for teaching academic writing in English: Corpus evidence. In S. Nuccorini (ed). *Phrases and phraseology: Data and descriptions*. Bern: Peter Lang: 85-105.

OAKEY, D.J. (2009). Fixed collocational patterns in isolexical and isotextual versions of a corpus. In P. Baker (ed.). *Contemporary Corpus Linguistics*. London: Continuum: 142-160.

SCOTT, M. (2008). *WordSmith Tools*. Liverpool: Lexical Analysis Software.

SINCLAIR, J.M. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

SINCLAIR, J.M. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.

SWALES, J.M. and FEAK, C. (2005). *English in today's research world*. Ann Arbor: The University of Michigan Press.

WRAY, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

The *Base lexicale du français*: a multi-purpose lexicographic tool[1]

Serge Verlinde[2]
K.U.Leuven

Abstract

The *Base lexicale du français* (BLF) is a freely accessible lexicographic tool for learning French vocabulary developed especially for the web. It groups a large number of functions which provide quicker, more efficient and user-friendly access to very precise information. To enable access to this information a brand new interface based on users' needs has been implemented. In addition, the *BLF* also offers high-performance reading, translation and writing assistants.

Keywords: lexicography, dictionary, web, language learning, French.

1. Introduction

In his recent study, Tarp (2008: 123) suggests that a high-performance lexicographic reference tool should be a kind of search engine enabling queries on a lexical database or on the web. This tool, called *leximat*, should allow any user with a particular communicative or cognitive need to access lexicographic data. The resulting data should then meet his needs.

Access to the content of electronic dictionaries, which are often based on previous paper versions, has improved considerably thanks to powerful search functions on pronunciation or fuzzy search with approximate string matching. Some dictionaries also allow plain text search or the use of regular expressions as on the *OPUS* website. Others such as the *CNRTL* (Centre National de Ressources Textuelles et Lexicales) website combine various lexical resources such as a monolingual dictionary, a dictionary of synonyms or a corpus. However, despite all improvements these sophisticated resources cannot always provide a quick answer to frequent questions raised by language learners (*e.g.* the correct spelling of words with irregular morphology or irregular verb forms) (Pruvost 2003). Furthermore, like consulting a paper dictionary, consulting an electronic dictionary requires specific skills that many users do not have (Miller 2006).

[1] This article is a translation of an article in French published in the proceedings of the Journées scientifiques "Lexicologie, terminologie, traduction" (LTT) (Lisbon 2009). I would like to thank Herlinda Vekemans for proofreading this translation.

[2] Leuven Language Institute, K.U.Leuven, serge.verlinde@ilt.kuleuven.be

This paper describes the design of a new lexicographic tool which provides quicker, more efficient and user-friendly access to the requested information, depending on the user's needs. We will use the new interface of the *Base lexicale du francais* (BLF) as an example. This is a freely accessible online dictionary meant for intermediate and advanced learners of French.

2. From communicative and cognitive needs to user interface design

Several possible queries of dictionary users have already been listed by Hausmann (1977: 144) and more recently by Tarp (2008: 77). Taking these lists as a starting point, we believe six types of consultation may be distinguished when a dictionary is used to solve a problem concerning a word, a word combination (multiword expressions) or the lexicon in general:

1. Information retrieval: for information on a word or word combination (*e.g.* gender, spelling, meaning, syntax)
2. Translation retrieval: to translate single words and multiword expressions to a foreign language
3. Verification function: to verify the use of a word or multiword expression or to check a possible translation
4. Assistant function: to provide contextual information while reading, translating or writing
5. Learning function: to discover how the lexicon is structured and avoid common errors
6. Exercises: to practise

Information and translation retrieval clearly meet communicative needs. Whereas the learning function and the exercises deal with cognitive needs, the verification function and the assistant function meet both communicative and cognitive needs, depending on the type of situation. Somebody may want to check the translation of a word because he needs it in a text (communicative need) or simply because he just wants to verify something he heard (cognitive need).

Merging these functions in one single application has significant advantages compared to the current dispersal of lexicographic resources. Moreover, being able to process an entire text by adding lexicographic information automatically to all the words (assistant function) saves the user a considerable amount of time. Information is contextualized and word by word consultation of the dictionary is no longer necessary.

The type of dictionary consultation determines the access to the lexicographic data. The homepage of the *BLF* reflects the six options listed above (*cf.* Figure 1).

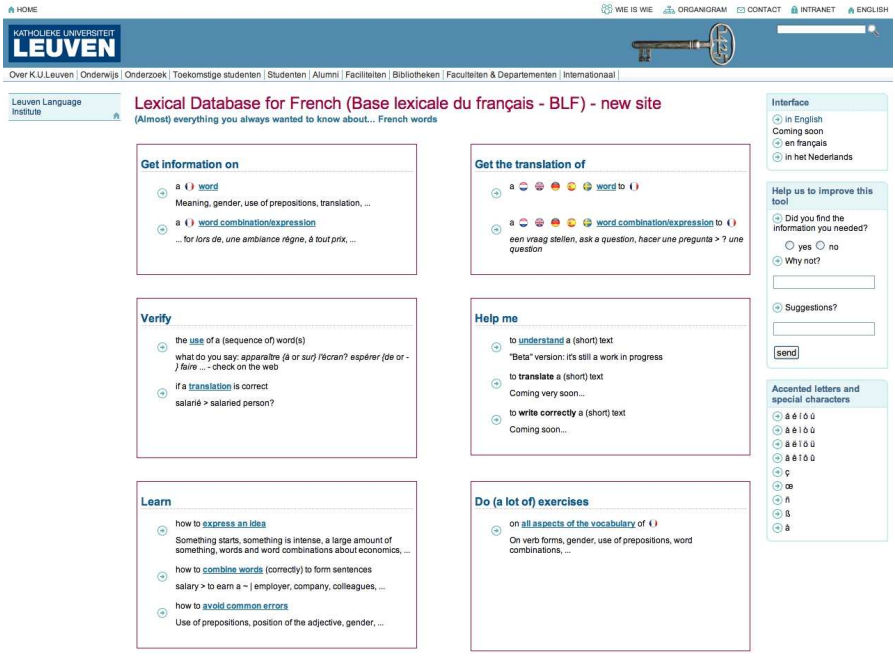


Figure 1. BLF homepage

Figure 2. Get information on a French word

Each access path (*e.g.* “Get information on”, “Verify”, Learn”) shows other query possibilities. Figure 2 lists the questions a user may submit to the database by selecting the option “Get information on a French word”.

The menus displayed allow users to:

- enter a word or multiword expression and click on the button corresponding to the user’s question. In answer to this query, the lexical database returns precise information, *e.g.* the gender of a submitted word, its meaning, its syntax;
- click on a button in order to extract structured information from the lexical database, *e.g.* a list of adjectives expressing intensification, a list of prepositional phrases;
- paste an entire text in a text box in order to obtain contextualized information, translation or writing aid.

As shown in Figures 3 and 4, information provided may take several forms.

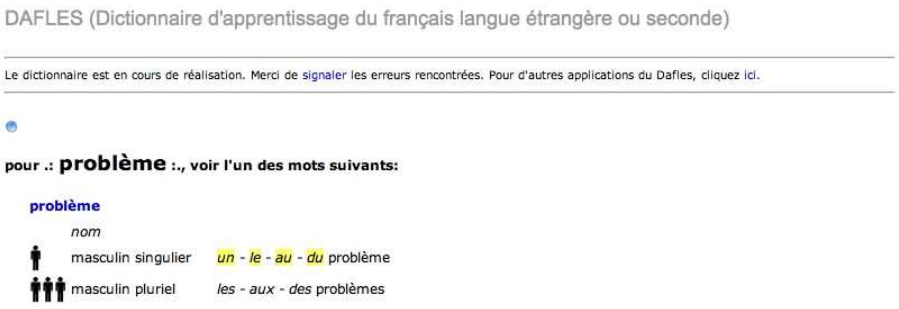


Figure 3. Gender of the word problème

Figure 4. Reading assistant

Figure 4 provides an example of the contextualized information a user receives when he submits an entire text to the reading assistant (see also section 5).

The architecture of the *BLF* provides significant improvement in accessing lexicographic data. The sole access by a text box or a word list as we find it in both paper and electronic dictionaries has been replaced by a user-oriented interface with multiple access paths leading to very specific information instead of a block of text. However, if the user requires more elaborate information, hyperlinks on the words displayed allow him to access the full *BFL* articles.

3. From interface to multivariate lexicographic description

According to Gouws (2007: 66), lexicographers tend to overestimate the linguistic competences of dictionary users. Language is a truly complex system and lexicographic description will thus be complex too. However, one has to admit that the overuse of metalanguage or abbreviations in dictionaries makes this lexicographic description less transparent. Therefore we opted, whenever possible, for a more didactic and user-friendly presentation. For example, we added articles to nouns in order to facilitate memorisation (*cf* . Figure 3).

In the learning function, a separate section is dedicated to a series of problems typically encountered by non-native speakers of French: position of the adjective, syntactic constructions for nouns, adjectives and verbs, etc. The *BLF* user will easily find this type of information concentrated on one page rather than scattered as in traditional dictionaries or only presented in outline form as in grammars.

The lexicographic description of the *BLF* also differs from the content of traditional learner's dictionaries for French (as a second or foreign language (*e.g.* Rey-Debove 1999). In the *BLF* a lot of attention is paid to encoding problems by providing the word profile of almost 13,000 words (Verlinde *et al*. 2006) and the presentation for any verb meaning and any verb complement (subject, direct object, etc.) of prototypical words and word combinations (Verlinde *et al*. 2004).

4. Use of web resources

The *BLF* describes the 6,500 most frequent lemmas of a newspaper corpus (Verlinde and Selva 2001). The lexicographic description of this set of words has not yet been fully completed. Fortunately, over the last years, numerous web resources for French have appeared (Habert 2005). In the *BLF*, we try to use these resources to fill the gaps of our own lexicographic description. Information from these sites is immediately accessible by shortcuts that directly open external sites on a specific page (*cf* . Figure 5 for the *problème* word family on the *Orthonet* website, starting from the query 'Words with the same meaning' as displayed on Figure 2).

Figure 5. Orthonet: the problème word family

External resources are also used to provide translations in the *BLF*. Apart from shortcuts to the websites of bi- or multilingual dictionaries, we also use freely available parallel corpora (*e.g.* the *OPUS* project, Tiedemann and Nygaard 2004), which are very interesting for the richness of translation equivalents found in their phrase-aligned texts.

Thanks to the integration of a multiplicity of information and resources, the *BLF* has become a very flexible and adaptive tool for users with very different levels of linguistic competence. Learners at an intermediate level who still have some doubts about the morphology of a word or the gender of a noun as well as linguists who want to carry out a detailed analysis of synonymy based on word profiles will find just the information they need.

5. Towards an integrated online help function

The online reading, translation and writing assistants are probably the most striking functions of the *BLF* (see Figure 4 above). The reading and translation assistant operate in a similar way: a submitted text is analysed and sent back to the user with information added to every word or multiword expression. By moving the mouse over a word or a multiword expression, this information is displayed in a pop-up window providing links to pages of the *BLF* (*e.g.* meaning, translation, synonyms).

The aim of the writing assistant, currently under construction, is to enhance users' awareness of common errors and to develop ways to avoid these errors by a systematic control procedure. First, the programme identifies syntactic and lexical patterns typical of frequent errors (position of the adjective, use of prepositions, use of *imparfait/passé composé*, etc.). Next, all occurrences of these patterns are grouped and displayed on the screen along with a small, didactic grammatical description in order to help to user to read his text over and to make necessary corrections to it.. Whenever possible, data from our lexical database are added to the feedback provided to the user (Verlinde *et al*. forthcoming).

The writing assistant differs from Word's grammar checker in that it does not correct errors in the text. Its aim is to enhance the skills needed for self-correcting and self-learning.

6. Future developments

A tracking and logging system registers all actions of the users on the *BLF* site by identifying more than 250 access paths to data from our website or to external sites. We expect that an analysis of these data will allow us to define the user's search behaviour more accurately and thus improve our lexicographic tool. Possible improvements could influence both the lexicographic content as well as the access paths to information.

We would also like to extend the interface by creating shortcuts to various resources for languages other than French. The interface could thus become a single access lexicographic portal site for any problem with the lexicon of many languages.

7. Conclusion

Lexicography can no longer be considered a craft industry (Rey 2008). It has definitively entered the internet age. Applying web database applications in a context of enriched lexicographic descriptions has opened new perspectives for the integration of resources and the development of new and efficient reading, translation and writing assistants and has in fact reshaped the very definition of 'dictionary'. All of this is in the obvious interest of a wide variety of users.

References

BLF (2009). *Base lexicale du français*. K.U.Leuven: Leuven Language Institute. ilt.kuleuven.be/blf.

CNRTL (2009). *Centre National de Ressources Textuelles et Lexicales*. CNRS: ATILF. www.cnrtl.fr.

GOUWS, R. (2007). Sublemmata or main lemmata: a critical look at the presentation of some macrostructural elements. In H. Gottlieb and J.E. Mogensen (eds). *Dictionary visions,*

research and practice. Selected papers from the 12th international symposium on lexicography, Copenhagen 2004. Amsterdam/Philadelphia: Benjamins: 55-70.

HABERT, B. (2005). *Instruments et ressources électroniques pour le français*. Gap/Paris: Ophrys. www.limsi.fr/Individu/habert/Publications/InstrumentsEtRessourcesElectroniquesPourLeFrancais.html.

HAUSMANN, F.-J. (1977). *Einführung in die Benutzung der neufranzösischen Wörterbücher*. Tübingen: Niemeyer.

MILLER, J. (2006). An investigation into the effect of English learners' dictionaries on international students' acquisition of the English article system. *International Education Journal*, 7(4): 435-445. ehlt.flinders.edu.au/education/iej/articles/v7n4/Miller/paper.pdf.

Opus (2009). *Opus – an open source parallel corpus*. urd.let.rug.nl/tiedeman/OPUS/.

Orthonet (2009). CILF. orthonet.sdv.fr/.

PRUVOST, J. (2003). *Some lexicographic concepts stemming from a French training in lexicology*. kdictionaries.com/newsletter/kdn11-03.html.

REY, A. (2008). *De l'artisanat des dictionnaires à une science du mot. Images et modèles*. Paris: Colin.

REY-DEBOVE, J. (1999). *Dictionnaire du français*. Paris: CLE International/Le Robert.

TARP, S. (2008). *Lexicography in the borderland between knowledge and non-knowledge*. Tübingen: Niemeyer.

TIEDEMANN, J. and NYGAARD, L. (2004). The OPUS corpus. In *Proceedings of the fourth international conference on language resources and evaluation (LREC'04). Lisbon, Portugal, May 26-28*. stp.ling.uu.se/~joerg/paper/opus_lrec04.pdf.

VERLINDE, S., PAULUSSEN, H., SLOOTMAEKERS A. and DE WACHTER, L. (forthcoming). La conception d'un didacticiel d'aide à la rédaction. *Revue française de linguistique appliquée*.

VERLINDE, S. and SELVA, T. (2001). Nomenclature de dictionnaire et analyse de corpus. *Cahiers de lexicologie*, 79: 113-139.

VERLINDE, S., SELVA, T., BINON, J. and PETIT, G. (2004). Les schémas actanciels dans le dictionnaire: point de convergence entre la morphologie et la sémantique lexicale. In G. WILLIAMS S. and VESSIER (eds). *Proceedings of the eleventh EURALEX international congress*. Lorient. II: 427-437.

VERLINDE, S., BINON, J. and SELVA, T. (2006). Corpus, collocations et dictionnaires d'apprentissage. *Langue française*, 150: 85-98.

Building an electronic combinatory dictionary as a writing aid tool for researchers in biology

Alexandra Volanschi[1], Natalie Kübler[1]
CLILLAC-ARP, Université Paris Diderot – Paris 7

Abstract

The present paper presents a methodology to explore the combinatorial properties of terms belonging to biology, based on the analysis of a large corpus of scientific articles. The research has led to the production of a writing aid tool meant to help non-native authors write scientific papers in English. The tool design is based on the evaluation of potential users' needs, which was undertaken by sending out a questionnaire to the teaching and research staff at the Life Sciences Department at the University Paris Diderot. The paper describes the corpus used, the first stage of data extraction, *i.e*. collocation extraction, as well as the specifications of the web interface allowing users to query data.

Keywords: phraseology, collocation extraction, electronic dictionary, writing aid, predicate-argument structures.

1. General background

It is an unquestionable fact that English is nowadays the "International Language of Science".[2] To take a single example, 95% of all publications indexed by the Science Citation Index are written in English. This reality has far-reaching consequences on the research activity worldwide. A researcher's activity is assessed in terms of the *impact factor* of his publications. As articles written in English have a wider readership, they have more chances of being cited and therefore, the journals publishing in English have a higher *impact factor*.

The "*anglophone grip*" (Swales 1990: 97) on published research communication is not uncontroversial. While for some English is a *lingua franca*, which has made possible scientific progress, for others, the dominance of English in published research is a harmful phenomenon, a *Tyrannosaurus Rex* (Swales 1997) both because it has led certain users – that is native speakers of English – to enjoy a preferred status and because it may lead to the "loss of specialised registers in otherwise healthy languages" (Swales 1997: 376).

[1] {avolansk,nkubler}@eila.univ-paris-diderot.fr

[2] The phrase was coined in the late 60s: English - An International Language for Science, Current Contents, 1967, vol. 1: 19-20.

We found that, in the field we worked on (biology) and judging by the journals on which our research is based (*cf.* Section 3.1), both these reasons for concern are justified. On the one hand, the statistical analysis of the corpora used in this study shows that

> "authors based in the Inner Circle in general and those based in the United States in particular, enjoy a disproportionately large percentage of publications and are more likely to be gatekeepers of published works" (Tardy 2004: 248).

On the other hand, Swales' concern over the loss of specialised genres seems to be justified as regards the use of French in scientific written communication. In the field of yeast biology for instance, the only written texts we could gather in French were popularization journal articles, teaching materials and, above all, PhD dissertations. The French law imposes PhD dissertations to be written in French. Advanced, leading edge research articles seem to be written in English exclusively.

The "*publish or perish*" imperative which governs a researcher's career may be read today as "*publish in English or perish*". This imperative is rendered discouraging for young French researchers – who are constrained to publish in English as early as the post-doctoral level – by the lack of specialised dictionaries and teaching materials adapted for their needs. The research reported in this paper is aimed at designing a writing aid tool meant to meet the needs of young French researchers in biology, and more generally of non-native speakers of English. This was achieved by devising a hybrid collocation extraction method (combining dependency parsing and a number of statistical heuristics) which was applied to a specialised corpus of research articles. The remainder of the paper is organised as follows: section 2 deals with the user needs evaluation and tool design, section 3 describes the corpus used and the data extraction method; the first, tentative query interface is presented in section 4 together with the improvements planned; finally, section 5 anticipates future work.

2. User needs evaluation and tool design

In order to better estimate our potential users' needs, we sent out a questionnaire to the teaching and research members of the Life Sciences department at the University Paris Diderot. To our satisfaction, we received 56 answers out of a total of 300 potential participants, which is a reasonable turnout, showing that researchers were concerned about the issues raised by the questionnaire. The high turnout is also due to the way in which the questionnaire was organised. It consisted of 15 questions alone – and we estimated that they would take about 12 minutes to answer. The questionnaire was sent by email and required participants to answer by email, thus avoiding any additional effort of downloading, filling in, saving and attaching a file to another e-mail. Throughout the questionnaire we supplied examples to illustrate the notions we were trying to inquire about. We did not use linguistic notions such as "collocation", "predicate-argument structure", and "idiomatic expression" which would have been irrelevant for researchers in biology.

The analysis of the questionnaire results has shown that English is truly a working language for French researchers as 96% of all written productions are in English (the texts written in French are PhD dissertations and teaching materials). In 95% of the cases, articles are written directly in English; very few researchers resort to translators. However, barely half (53%) claim to be thinking in English when they write in English, which implies that the other half go through the mental act of translation.

Regarding writing aid needs, 90% of the participants to the questionnaire stated that they use other scientific articles as a writing assistance tool, looking mostly for scientific information (68%), but also for phraseological patterns, connectors, sentence adverbs (64%), and term phraseology (51%). Through our research, we aim to facilitate and systematize access to this kind of information. Not surprisingly, the difficulties researchers encounter are related to grammar (62.5%), knowledge of idiomatic expressions (70%), the influence of French in the writing process (64.28%) and, to a much lesser extent, knowledge of specialised terms (14%). While researchers acquire a thorough knowledge of specialised terms early in their careers by reviewing the literature in the field, what poses a problem is using these terms in context.

In order to evaluate the usefulness of the various solutions we were considering for the writing aid tool, we supplied a number of examples and found that participants to the questionnaire would be interested in terminological collocations (83%), "general scientific language" (Pecman 2004), collocations and idiomatic expressions (87%), verb argument structures (73%) and, to a lesser extent, concordances (55%). The last figure seems to suggest that users lacking linguistic training are unlikely to be able to generalize beyond a number of given examples. Of all participants, 55% would prefer a bilingual writing aid tool. However, given the scarcity of research articles written in French, we have chosen to develop a monolingual tool, at least in its initial stage.

Numerous studies, among which Nesselhauf (2005), Granger (1998), and Howarth (1996; 1998), have shown that mastering collocations is one of the greatest challenges in second language acquisition and analysed the typical errors non-native speakers make, such as literal translation from the mother tongue (L1), simplifying or overgeneralising the elements of the L2, etc. Nevertheless, mastering phraseology is one of the proofs of belonging to a discourse community; in particular, it is one of the proofs of belonging to a scientific community (Gledhill 2000). Both biology terms and cross-disciplinary scientific terms have a specific phraseological behaviour, which is codified and must be adopted in order to gain acceptance in the scientific community. The design of the writing aid tool is therefore centered on the exploration of the combinatory properties of specialised terms and terms belonging to general scientific language. We began by extracting from the corpus *restrictive collocations* (defined in section 3.2.1.), which are easier to describe formally and therefore easier to retrieve. The study will be extended in the future to the extraction and encoding of longer idiomatic expressions and formulae specific to scientific discourse, associated with the rhetorical function they serve.

3. Corpus and data extraction method

3.1. Corpus design

Two successive experiments were conducted for collocation extraction. In the first phase, we focused on the field of yeast biology. Yeasts are eukaryotic microorganisms used as a model organism to study certain biological processes in humans, such as cellular division, metabolic diseases (such as Friedrich's ataxia) or cellular regulation and deregulation. For this first corpus, called LangYeast, we selected research articles published over the last 10 years in two journals highly specialised in yeast research: *Yeast*[3] and *FEMS Yeast*[4] as well as a number of research articles from *Science Magazine*[5] dedicated to yeast research. While the former are highly specialised in yeast research (and therefore have a lower impact factor: 1.97 and 2.27), *Science Magazine* is a journal that covers the full range of scientific disciplines (and therefore reaches a larger audience; its impact factor in 2006 was of 30.02). We gathered 1461 articles from these journals totalling 5.5 million words. We have chosen these journals because they were available in the electronic holdings of our university library. Nevertheless, these texts are subject to copyright restrictions, which posed a problem as we wanted to supply plenty of examples to illustrate the collocations extracted from the corpus. This could be achieved by quoting very brief text fragments and formally acknowledging any material used. However, in order to avoid copyright issues we are now running a second experiment on freely accessible, open access material.

In the second experiment on collocation extraction, we used research articles from the PLoS journals[6]: *PLoS Biology* (5.5M words), *PLoS Pathogens* (3.8M words), *PLoS Genetics* (4.1M words), *PLoS Medicine* (2.2M words), *PLoS Computational Biology* (3.8M words). Given that the PLoS articles are at the same time peer-reviewed (the quality of the articles published is thus guaranteed) and freely available online, the PLOS journals have a quite high impact factor (ranging from 6.2 for *PLoS Computational Biology* to 13.5 for *PLoS Biology*).

In order to improve the accuracy of the results, we aimed to use only a subset of articles, namely those written by native speakers of English. However, as this information is virtually impossible to check, we used an approximation and selected articles having at least one author based in the United States, the United Kingdom or Australia. A statistical analysis of the corpora used has shown a significant correlation between these percentages and the journal's impact factor. While the percentage of articles having at least one author based in English-speaking countries is relatively low in highly specialised, low impact factor journals such as *Yeast* and *FEMS Yeast* (28% and 39%), these percentages are significantly higher in high impact factor journals

[3] Published by Wiley Interscience (http://www3.interscience.wiley.com).

[4] Published by Blackwell Synergy (http://www.blackwell-synergy.com) for the Federation of European Microbiological Societies (FEMS).

[5] http://www.sciencemag.org

[6] http://www.plos.org

(88% in *Science*), even approaching 100% in *PLoS Pathogens*, where this percentage is as high as 98.03%. Without trying to draw any firm conclusions based on these statistics, it would appear as though French scientists stand more chances of being published when collaborating with an author based in English-speaking countries, and more precisely in the USA.

If all authors published in the PLoS journals are considered together, more than half (54%) are based in the United States, 10.22% are based in the UK, and 1.8% in Australia: two thirds of the authors published in PLoS are based in English-speaking countries. Swales' reasons for concern seem to be justified.

3.2. Data extraction

3.2.1. A working definition of collocation

The vast body of literature on phraseology which has developed over the last fifty years abounds with definitions of collocations, each highlighting various defining features of collocations such as the number and nature of the elements making up the collocations, the syntactic relationship and the distance between the two, the frequency, their arbitrary nature and the compositionality of the combination, and the statistical association measure of the two (*cf.* Siepmann 2005; Bartsch 2004).

We have chosen a number of these defining features and proposed a working definition of collocations, well adapted for the purposes of this research. This definition may not take into account the semantic aspects of collocations such as compositionality, but only the formal aspects which can be implemented in an extraction tool. We defined collocations as binary polar recurrent combinations, the two elements of which are in a direct syntactic relation. The polarity or orientation of the base-collocate combination is parallel to the syntactic dependency relation. We computed an association measure (mutual information) of the base and the collocate. However, we decided against using this association measure in the second experiment on the PLoS corpus, mainly because we consider it superfluous: given the syntactic dependency between the two co-occurrents, it is natural to assume that they are associated. Moreover, we found that using the strength of association as a sorting criterion of the collocates is less pertinent than sorting by frequency; we have therefore decided to implement only a minimum frequency threshold of 3 occurrences.

By choosing syntactic dependency as a defining criterion we have opted for *relational co-occurences* (Evert 2005: 19), that is co-occurrences in a syntactically defined window, as opposed to *positional co-occurences, i.e.* co-occurrences within a n-word window. Relational co-occurrences have been criticised for imposing notions of a given linguistic theory on the data (Sinclair 1991) and also for the fact that automatic linguistic analysis unavoidably introduces a number of errors. We have nevertheless chosen to extract relational co-occurrences, because they are distributed in homogeneous, easier to interpret classes.

3.2.2. Collocation extraction

The corpus was analysed using a classical sequence of processing stages: sentence splitting, tokenisation, part of speech tagging, lemmatisation and – the key stage in the process – dependency parsing. For the first stages we used a number of in-house tools and resources which take into account the specificities of the Language for Specific Purposes (LSP) we were working on (tokenisation rules, specialised vocabulary, etc.). For the dependency parsing we used the Stanford parser[7], which produces 48 typed dependencies the choice of which was "motivated by practical rather than theoretical concerns" (Marneffe *et al* 2006: 449). For each sentence in the corpus, the parser produces a list of typed dependencies – as illustrated here for *These vectors transcribe two genes simultaneously* – from which relational co-occurrences are extracted:

det(vectors-2, These-1)
nsubj(transcribe-3, vectors-2)
num(genes-5, two-4)
dobj(transcribe-3, genes-5)
advmod(transcribe-3, simultaneously-6)

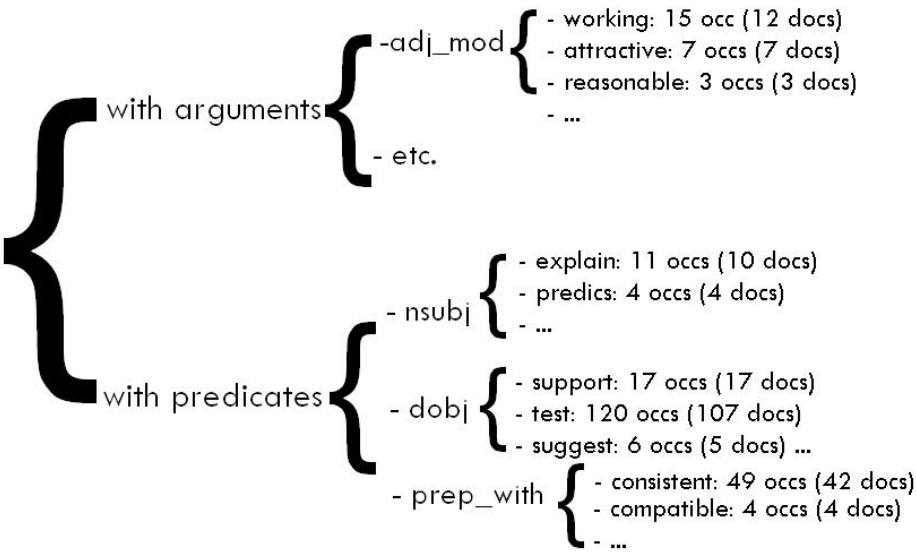


Figure 1. Collocation extraction for the noun hypothesis

Starting from the dependency parsed corpus, collocation extraction is performed along the lines suggested by Lin (1998), then implemented in WordSketch for English (Kilgarriff and Tugwell 2001) and Les Voisins de le Monde for French (Bourigault

[7] http://nlp.stanford.edu/downloads/lex-parser.shtml

and Galy 2005). It consists in the extraction of recurrent typed dependencies (as illustrated in Figure 1 for the noun *hypothesis*), recording each time the number of occurrences, and the number of different documents in which the co-occurrence was found. For both these measures, we set a minimum threshold of 3.

3.2.3. Predicate-argument structure extraction

By examining the results of the collocation extraction phase, we noticed one limitation of this approach. We were able for instance to extract from the corpus recurrent co-occurrences of the verb *to investigate* with modifying adverbs (*e.g. systematically, thoroughly, fully*), with direct objects (*e.g. the effect, the role*), with *in* prepositional objects (*e.g. in species, in mice, in strains*) because they had more than 3 occurrences. However, we were not able to extract any dependency with a *for* prepositional object (*e.g. for properties, for characteristics, for features*). Even though we extracted 60 *for object* dependency relationships attached to the verb *to investigate*, each object in particular had only one or two occurrences in the corpus, thus falling under the minimum frequency threshold which we set to 3.

Particular relational co-occurrences might be recovered by increasing the corpus size, but we suggest that better results might be obtained by looking not only for recurrent typed dependencies *base – typed dependency – collocate*, but also for recurrent dependencies of a base and a **collocate type** *base – typed dependency – collocate class*. We extracted from the dependency parsed sentences all dependencies of verbs in the corpus, analysed these manually and proposed *predicate-argument structures* or collocational schemas for a number of verbs. At the interface between syntax and semantics, these structures describe the semantic arguments and their possible syntactic realisations. We believe this type of analysis will allow users to gain a clearer picture of verb usage, which relational co-occurrences represent only partially.

Providing examples of verb predicate-argument structures is particularly useful in the case of specialised verbs: a lot of verbs in biology are formed by metaphorical extension from language for general purposes verbs: *to express, to translate, to transcribe, to block, etc*. However, these verbs develop a completely different argument structure in language for specific purposes. As illustrated in Table 1 by (part of) the predicate-argument structure of the verb *to transcribe*, we would suggest that, apart from the initial metaphorical link, the verb has nothing in common with the general language verb *to transcribe*, which has a very simple predicate-argument structure: Arg0:copier, agent, Arg 1:thing copied.[8] A general language dictionary would prove useless as a writing aid for verbs of this kind.

[8] UNIFIED VERB INDEX: http://verbs.colorado.edu/verb-index/index.php (the union of VerbNet, PropBank, FrameNet and OntoNotes Sense Groupings)

Argument	Example
Arg 0: agent/ sujet	RNAp, RNA polymerase, RNAp III
Arg1: entity transcribed/object	gene, genome, ortholog, repeats, etc.
Arg2: entity after transciption/*into* object	cDNA, mRNA,
Arg3: location, organ/tissue expressing transcription	**in** tissue
Arg 4: direction	**towards** the telomere/centromere, **in** the same/opposite **direction**
Arg 5: source/transcription site	**from** promoter region
Arg6: conditions	**under** non-stress conditions, **in response to** heat shock, **in the presence/absence** of carbon source
ArgMod	divergently, convergently, constitutively, highly/at a high level, strongly, periodically

Table 1. Proposal of predicate-argument structure of the verb to transcribe

Predicate-argument structures could not reliably be constructed automatically and need to be validated by an expert in the field. We have manually built 20 such predicate-argument structures and intend to include them in the query interface, described in the following section.

4. The query interface

A vast body of in-depth studies has been devoted to the encoding of collocations, and more precisely in terminological databases and dictionaries. Examples include Benson *et al*. (1986), Mel'cuk *et al*. (1995), Heid and Freibott (1991), Heid (1992), Béjoint and Thoiron (1992), Cohen (1986), Meynard (1997), and Binon *et al*. (2000). Most of these dictionaries or databases were built manually, and therefore afford a much more fine-grained encoding of collocations. They do not, however, address a readership lacking linguistic background. While trying to retain the most interesting proposals these studies make (such as the idea of encoding collocations both under the base and collocate entries), we have devised a simple collocation encoding system and a minimalist query interface.

The data extracted may be queried via the first version of the search interface.[9] As represented in Figure 2 by results found for the verb *to investigate*, this version

[9] Hosted on the Institute Jacques Monod website: http://ytat2.ijm.univ-paris-diderot.fr/LangYeast/

supplies co-occurrences of the search term with collocates linked to it by a syntactic relationship, as well as statistical information such as the frequency of the co-occurrence, its coverage – measured by the number of different documents in which it appeared, and mutual information. For each co-occurrence, the user is provided with one example chosen randomly from the corpus, which is displayed in a pop-up window.

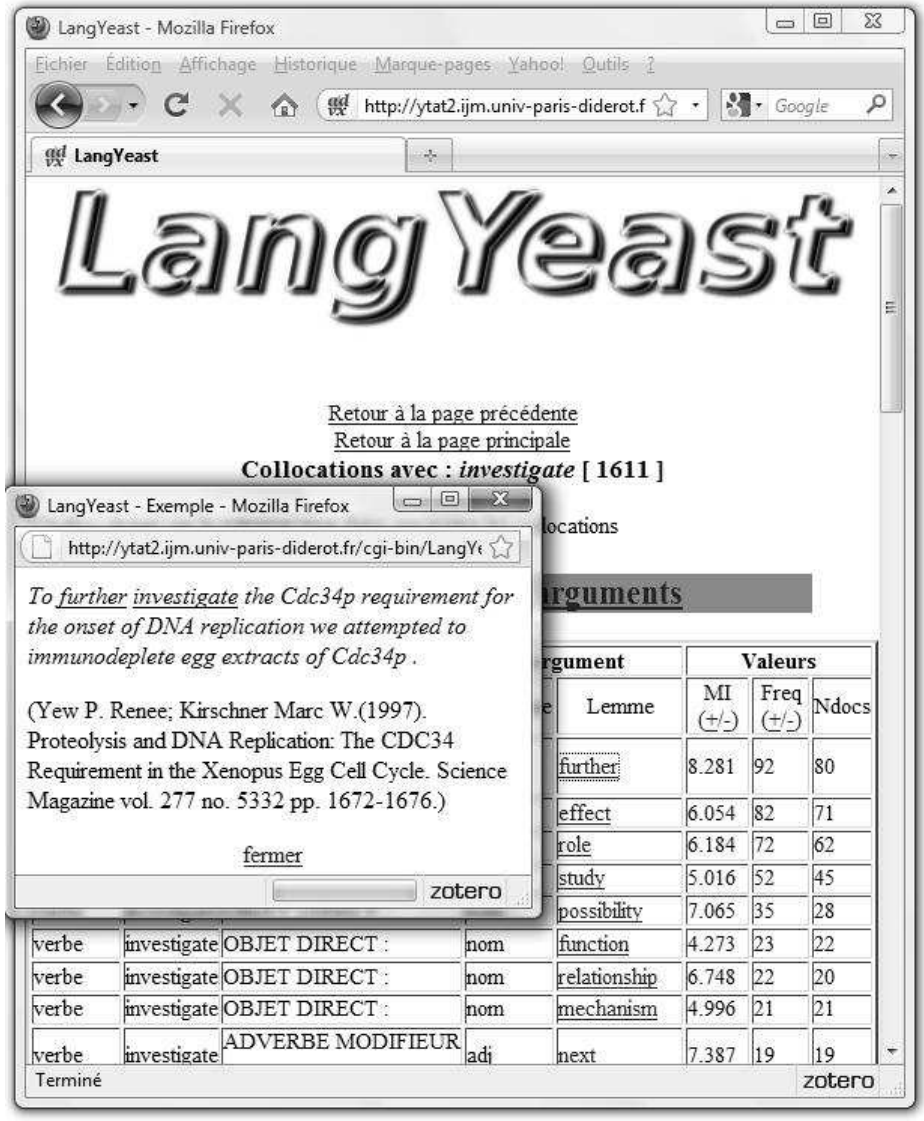


Figure 2. First version of the combinatory dictionary query interface

5. Future work

Several procedural improvements of the collocation extraction process and of the query interface will be implemented. In its current stage, the dictionary contains 5,200 entries (selected by examining frequency lists extracted from the corpus) and 77,000 collocations. Obviously, the results contain a lot of noise and should ideally be manually validated. Prior to this validation stage, two means of reducing the number of collocations have been envisaged. On the one hand, a lot of collocations are actually complex terms; for the second experiment on the PLoS corpus, we plan a complex term extraction stage prior to dependency parsing. On the other hand, a lot of collocations with instances of gene, bacteria, protein or yeast names could be grouped under collocations with semantic classes, defining the term's semantic preference. Thus, collocations with particular instances such as *yfh1*, *E. coli*, and *S. cerevisiae* could be replaced by collocations with the corresponding semantic classes {*gene*}, {*bacteria*}, {*yeast*}.

The query interface reflects the process through which data was extracted. However, it is not very appropriate for the intended users. Indeed, for biologists, notions such as *predicate, argument, verb, modifying adverb* are completely irrelevant. In the second version of the query interface we intend to simply group together collocates linked to a base by a given syntactic relation, without necessarily supplying a linguistic name for each class; a generic name such as 'type 1' could be used and associated with the explicit, extensive list of collocations: *to test a hypothesis, to confirm a hypothesis, to refute a hypothesis*, etc. We think that statistical information such as frequency or coverage would be interesting to keep as well. The second version of the query interface should also supply an access to examples of predicate-argument structures as well as longer idiomatic formulas belonging to general scientific language which we intend to extract. While access to collocations and predicate-argument structures will be given via a search term, as in the current version of the query interface, idiomatic formulas will be grouped and accessed via the rhetorical role they play. For instance, formulas such as *it is commonly/generally/universally/widely accepted that, it has been often asserted/noted/claimed/argued that*, would be grouped under "Making topic generalisations".

In addition to these methodological improvements, we intend to formally test the efficiency of the writing aid tool and to create teaching material derived from the corpus and the combinatory dictionary developed.

References

BARTSCH, S. (2004). *Structural and functional properties of collocations in English. A corpus study of lexical and pragmatic constraints on lexical co-occurrence.* Tübingen: Verlag Gunter Narr.

BEJOINT, H. and THOIRON, P. (1992). Macrostructure et microstructure dans un dictionnaire de collocations en langue de spécialité, *Terminologie & Traduction, Office des publications officielles des Communautés européennes*. Luxembourg: 513-522.

BENSON, M., BENSON, E. and ILSON, R. (1986). *The BBI Combinatory Dictionary of English*. Amsterdam: John Benjamins.

BINON, J, VERLINDE, S., VAN DYCK, J. and BERTELS, A. (2000). *Dictionnaire d'apprentissage du français des affaires*. Paris, Didier.

BOURIGAULT, D. and GALY, E. (2005). Analyse distributionnelle de corpus de langue générale et synonymie. In *4e Journées de la linguistique de corpus*. Lorient: 163-174.

COHEN, B. (1986). *Lexique de cooccurrents. Bourse – conjoncture économique*. Montréal: Linguatech.

EVERT, S. (2004, published 2005). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

GLEDHILL, C.J. (2000) *Collocations in Science Writing*. Tuebingen: Gunter Narr Verlag (*Language in performance*, 22).

GRANGER, S. (1998). Prefabricated patterns in advanced EFL writing: collocations and formulae. In A.P. Cowie (ed.). *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press: 145-160.

HEID, U. and FREIBOTT, G. (1991). Collocations dans une base de données terminologique et lexicale, *Meta*, XXXVI(1): 77-91.

HEID, U. (1992). Décrire les collocations. Deux approches lexicographiques et leur application dans un outil informatisé. In *Terminologie & Traduction, Office des publications officielles des Communautés européennes*, Luxembourg: 523-548.

HOWARTH, P. (1996). *Phraseology in English Academic Writing: Some Implications for Language Learning and Dictionary Making*. Tübingen: Max Niemeyer.

HOWARTH, P. (1998). The Phraseology of Learners' Academic Writing. In A.P. Cowie (ed.). *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press: 161-186.

KILGARRIFF, A. and TUGWELL, D. (2001). WORD SKETCH: Extraction, Combination and Display of Significant Collocations for Lexicography, In *Proceedings of the Workshop on Collocations: Computational Extraction, Analysis and Exploitation*, ACL-EACL 2001. Toulouse: 32-38.

LIN, D. (1998). Extracting Collocations from Text Corpora, First Workshop on Computational Terminology, In *COLING-ACL '98*. Montréal: 57-63.

DE MARNEFFE, M-C., MACCARTNEY, B. and MANNING, C.D. (2006). Generating Typed Dependency Parses from Phrase Structure Parses. In *LREC* 2006: 449-454.

MEYNARD, I. (1997). Approche hypertextuelle via HTML pour un outil de consignation bilingue des combinaisons lexicales spécialisées. In *Actes du Congrès international de terminologie*. San Sebastian: 675-689.

MEL'ČUK, I., CLAS, A. and POLGUERE, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve: Duculot.

NESSELHAUF, N. (2005). *Collocations in a Learner Corpus*. John Benjamins Publishing Company.

PECMAN, M. (2004). *Phraséologie contrastive anglais-français : analyse et traitement en vue de l'aide à la rédaction scientifique*. PhD dissertation, University of Nice.

SIEPMANN, D. (2005). Collocation, Colligation, and Encoding Dictionaries. Part I: Lexicological Aspects. *International Journal of Lexicography*, 18(4): 409-443.

SINCLAIR, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

SWALES, J. (1990). *Genre Analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

SWALES, J. (1997). English as a Tyrannosaurus rex. *World Englishes*, 16(3): 373-382.

TARDY, C. (2004). The role of English in scientific communication: lingua franca of Tyrannosaurus rex? *Journal of English for Academic Purposes*, 3(3): 247-269.

Are vector-based approaches a feasible solution to the "tip-of-the-tongue" problem?

Michael Zock[1], Tonio Wandmacher[2], Ekaterina Ovchinnikova[3]
LIF-CNRS, CEA-LIST, Institute of Cognitive Science

Abstract

Our goal is to help people find the word they are looking for when producing language (writing, speaking), a communication-mode often neglected by dictionary builders. Word finding problems manifest themselves in various forms via delays and quality of output, ranging from incorrect production (wrong term), to aborted productions (tip-of-the-tongue problem) or complete silence. Obviously, failures can be relative, and they may have various causes: ignorance, lack of usage or low frequency of the target word, semantic or phonological similarities, *i.e.* interference, etc.

We will be concerned here only with the "tip-of-the-tongue" problem, a state characterized by the fact that an author fails to produce a word even though s/he knows it. The word has been memorized, but is momentarily inaccessible. While one can think of various strategies or tools to help authors to overcome this state, we will consider here only one of them, LSA (Latent Semantic Analysis), to see how well it is suited to achieve our goal.

Keywords: lexical access, association, tip-of the tongue problem, vector-based methods, LSA.

1. The problem

Language production requires choosing words, which implies search in the lexicon. People usually start by consulting their mental lexicon (brain) before resorting to an external resource, something they do only in case of failure, dissatisfaction, or if they are particularly motivated and have time. Fortunately, generally this is not needed, as search is successful. Words are retrieved quickly and apparently without effort. Nevertheless, there are cases where things go wrong: one does not know the target word, or, a problem we are interested in here, one fails to access the word even though one knows it. This is known as the tip-of-the-tongue (TOT) problem, a consciousness state perceived as a knowledge gap, with the target word being imminent but elusive, despite very active search (James 1890). That the author knows (*i.e.* has memorized) the word, can easily be shown. S/he has produced the word (possibly in the (recent)

[1] LIF-CNRS, Université de la Méditerranée, Marseille, France, michael.zock@lif.univ-mrs.fr
[2] CEA LIST, Fontenay-aux-Roses, France, tonio.wandmacher@cea.fr
[3] Institute of Cognitive Science, University of Osnabrück, Germany, eovchinn@uos.de

past, s/he can recognize it immediately and without any mistake if it is presented in a list, and s/he may be able to produce it later on.

People experiencing the TOT state have a strong **f**eeling **o**f **k**nowing (FOK) that the retrieval of the target word is within reach (Brown 1991; Smith 1994). This feeling of imminence can be sustained by various facts. People feel that they know the word, and that it is about to pop up, ready to cross their lips at any moment. Next to conceptual and etymological knowledge (origin: Greek or Latin), they seem to know many other aspects concerning the target word: its number of *syllables*, *prosody*, *beginning* / *ending* (Brown and Mc Neill 1996), *part of speech* (noun, verb, adjective, etc.), and even *gender* (Vigliocco *et al.* 1997). Strangely enough, the resolution of the problem occurs spontaneously, possibly at a time when deliberate retrieval attempts have been abandoned (Burke *et al*. 1991; Reason and Lucas 1984).

2. Some psychological models accounting for lexical access

Studying speakers' performance, Fay and Cutler (1977) observed that people tend to make two kinds of errors: meaning-based substitutions (*left* instead of *right*) or substitutions based on similarity of form (*historical* instead of *hysterical*). Given the little evidence of interference between these two components, they claimed lexical access to be a sequential, *feed forward*-process, meaning choices taking place before computation of form, the latter having hardly ever an effect on lemma choices (feeding back).[4] This view is shared by many psychologists (*e.g.*, Bock and Levelt 1994)[5], according to whom the process takes place in the following way: given some information (semantic, conceptual), a *lemma* is retrieved, triggering the activation (or computation) of a *lexeme*, the lemma's corresponding phonological, or graphemic form.[6] For a more detailed and sophisticated account see Levelt *et al*. (1999).

The two-stage model, based on speech error data (Fromkin 1973, Fay and Cutler 1977) and studies of the TOT-problem, is widely accepted (Butterworth 1982; Kempen and Huijbers 1983). The major differences concern the relative autonomy of the modules and the direction of the information flow (uni- vs. bi-directional; *cf.* Caramazza 1997; Dell 1986; Levelt *et al*. 1999). That words may be accessed on the basis of sounds may not be straightforward, but has been shown in a classic experiment by Brown and McNeill (1966). Knowing that TOT states tend to occur with low-frequency words, they presented their definitions, asking subjects to produce the corresponding word.

[4] Whether there is feedback between these two components remains an open issue and shall not concern us here.

[5] But see Caramazza (1997), Dell *et al*. (1999), Badecker *et al*. (1995), who have different views concerning information flow and the lemma/lexeme distinction.

[6] Note that the term *lemma* here has a different meaning from the one it usually has in computational lexicography: content and form are considered separately. For psychologists a *lemma* is an abstract entity containing semantic and syntactic information (part of speech), but devoid of a phonological form (phonemes, syllabic structure, intonation curve). This information is encoded with the *lexeme*.

Interestingly, having received the definition of, let's say, *sampan*, people were more inclined to produce *sarong*, being semantically completely unrelated, than a related word like "yacht" or "houseboat".

This is how things are said to go if everything works fine, yet, as we all know, this is not always the case. According to the two-stage-model, lexical access can be hampered at either level, *i.e.* during lemma retrieval or during the computation of the phonological form (lexeme). While in both cases we are having a *word finding problem*, only in the second case, we speak of a *TOT-problem*. For a discussion concerning possible explanations of this kind of delay or failure, see Béroule and Zock (2009).

3. Our goal: assist the language producer

As we all know, in case of word finding problems one typical strategy consists in resorting to an external resource (dictionary). Dictionary users typically pursue one of two goals: (a) as a *decoder* (reading, listening), they are looking for the definition or translation of a specific target word, while (b) as an *encoder* (speaker, writer) they aim to find a lexical item expressing a concept and fitting into a given sentence slot (frame). We will be concerned here with the encoder's perspective.

While most dictionaries are better suited to assist the language receiver than the text producer, efforts have been made to improve the situation. Actually, *onomasiological* dictionaries are not new at all. Some attempts go back to the middle of the 19th century. The best known is, beyond doubt, Roget's *Thesaurus* (Roget 1852), but there are also T'ong's *Chinese and English instructor* (T'ong 1862), Boissiere's and Robert's *analogical dictionaries* (Boissière 1862; Robert *et al.* 1993), to name just those. Newer work includes *Longman's Language Activator* (Summers 1993) and various network-based dictionaries: *WordNet* (Fellbaum 1998), *MindNet* (Richardson *et al*. 1998), *HowNet* (Dong and Dong 2006) and *Pathfinder* (Schvaneveldt 1989). There are also proposals by Fontenelle (1997), Sierra (2000), Moerdijk *et al*. (2008), diverse *collocation dictionaries* (*e.g.* the BBI Dictionary of English Collocations, the Oxford Collocations Dictionary for Students of English), Bernstein's *Reverse Dictionary* and Rundell and Fox's (2002) *MEDAL*, a hybrid version of a dictionary and a thesaurus, produced with the help of Kilgarriff's Sketch Engine (Kilgarriff *et al*. 2004). While, obviously, a lot of progress has been made, we believe that more can be done.

As mentioned already, psychologists have shown that speakers being in the TOT-state have partial knowledge concerning the lexeme they intend to produce. We would like to use this information, no matter how poor it may be (partial or imperfect input), in order to help the authors to find the word they are looking for. In other words, given partial input we will try to guide their navigation, providing hints to lead them towards the target word.

To achieve this goal, Zock and Schwab (2008) have proposed to enhance an existing electronic dictionary by adding an index based on the notion of association. Their idea is basically the following: mine a well balanced digital corpus in order to capture the target user's world knowledge and construct a huge association matrix. The latter contains on one axis the *target words* (the words an author is looking for, *e.g.* '*fawn*') and on the other the *trigger words* (words likely to evoke the target word, *e.g.* '*young'*, '*deer'*, '*doe'*, '*child'*, '*Bambi'* etc.). At the intersection, they suggested to put frequencies and the type of link holding between the trigger- and the target-word (*e.g.* '*fawn--isa_a--deer*').

Once this resource is built, search is quite straightforward. The user provides as input all the words coming to his/her mind when thinking of a given idea or a lexicalized concept, and the system will display all connected, *i.e.* associated, words. If the user can find the item he or she is looking for in this list, search stops, otherwise it will continue (indirect associations requiring navigation), the user giving another word, or using one of the words contained in the list to expand the search space.

Remains, of course, the question of how to build this resource, in particular, how to populate the axis devoted to the *trigger words*, *i.e.* access keys. While Zock and Schwab (2008) use direct co-occurrence measures (1st order approaches) to determine association, there has been work suggesting that 2nd order approaches, *i.e.* vector-based models, are better suited to this end.[7] One of the main goals in this work is to verify if this is the case.

4. Vector-based approach

4.1. Latent Semantic Analysis, a second order approach

While Latent Semantic Analysis (LSA, Deerwester *et al.* 1990, Landauer *et al.* 1998) has initially been developed for improving information retrieval, it has been used for many tasks[8] since, and sometimes with remarkable success. For example, LSA was able to retrieve relevant documents even if they did not share a single element with the query. Hence the claim that LSA could reveal semantic relatedness on the basis of latent, hidden, *i.e.* indirect, information. If this is really so, then LSA should be a candidate for building a tool helping people to overcome the TOT state, a hypothesis we try to verify in this work. In the remainder of the paper we describe *Word Finder*, a prototypical system making use of this approach, and we present a small experiment in order to evaluate its capacities to resolve TOT states.

[7] Word space models like Latent Semantic Analysis (LSA) (Dumais 1990), or Hyperspace Analogue to Language (HAL) (Lund and Burgess 1996) are typical representatives of this approach.

[8] Automatic evaluation of student essays (Landauer *et al*. 1998), knowledge acquisition (Landauer and Dumais 1997), automated summarization (Wade-Stein and Kintsch 2003), automatic hyponymy extraction (Cederberg and Widdows 2003), etc.

4.2. Word Finder

Word Finder (WF) is supposed to help authors being in the TOT-state to find the target word. To this end the author provides all the information he can access at present, and the system returns a ranked list of candidate words among which the target word is supposed to be found.

4.2.1. Information flow

A typical search proceeds as follows. The user specifies a query (input) in the WF interface. The query is then forwarded to a standard web search engine (*e.g.* Google), returning a list of n top-ranked document snippets (example phrases). From these snippets WF extracts a set of candidate terms which are then ranked according to their semantic similarity with the query. The ranked list is finally returned to the user.

4.2.2. System interface

Figure 1 shows the prototypical interface of Word Finder. It comprises a field containing the query (here, "*word or sentence that reads the same backward and forward*"), the *output* box (ranked list of answers: 'palindrome', 'verse', 'like', etc.), as well a number of control panels: the words' *relative importance* (tf-idf weight) with respect to a *reference corpus*, the number of snippets to take into account and *its semantic relatedness* with the query (LSA parameter).

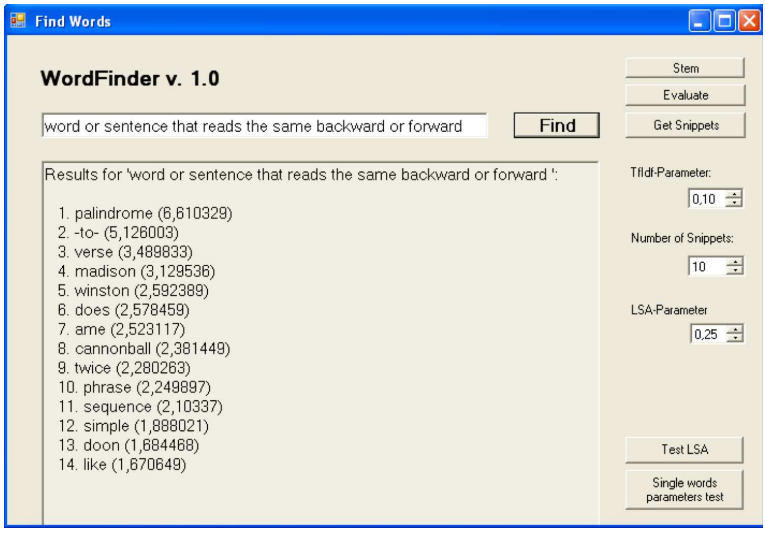


Figure 1. Word Finder interface

4.2.3. WF engine

In order to compute the output to a given query, *i.e.* return relevant terms, WF performs a series of operations on the initial document snippets. More precisely, it

creates a list of candidates, composed of all the words contained in the snippets (non-content (*i.e.* 'stop') words are filtered out). Next it applies a stemmer to the user query and initial candidate list. Finally, the stemmed candidates are ranked using the tf-idf and LSA measure.

4.2.3.1. Tf-idf measure

The tf-idf measure (term frequency-inverse document frequency) is an indicator of a word's specificity, which is an important feature for identifying a TOT word. For a word *w* it is based on the following equation: tf-idf(*w*) = *Df*(*w*) / log(*BaseF*(*w*)), where *Df*(*w*) and *BaseF*(*w*) refer respectively to the number of snippets containing *w* and to *w*'s frequency in a reference corpus. In our case, *ukWac*, a corpus of English web documents containing about 2GB of words, was used for determining the base frequencies.

4.2.3.2. LSA measure

Using distributional similarities of words, LSA builds a *semantic space* in which every term (word or expression) is represented as a vector. These vectors can then be compared to one another via vector similarity measures (*e.g*. cosine).

To compute LSA semantic relatedness we have used a newspaper corpus of 108 million words (*The Times* and *The Guardian* of the years 1996 and 1998). We calculated the LSA word space using the Infomap toolkit.[9] The initial co-occurrence matrix (window size: +-75 words) comprised 80,000×3,000 terms; it was then reduced to 300 dimensions by using SVD (singular value decomposition). Since the words of the query and those contained in the document were both represented as vectors, we could compute the semantic relatedness between any two of them (and even sets of words) as a cosine between corresponding vectors. The *LSA* measure generally gives reliable results concerning semantic relatedness of words, provided that there is enough evidence in the reference corpus.

4.2.3.3. Ranking

The ranking measure used for ordering the candidate list is a linear interpolation of the *tf-idf* and *LSA* measures described above. Thus, the rank *rm* of the word *w* is computed as follows: *rm = Df / Tf-Idfweight + LSA-weight * LSAval*, where *Df* refers to the number of snippets containing *w*, *BaseF* is *w*'s frequency in the reference corpus; *LSAval* expresses the LSA similarity of *w* with respect to the query words (cosine); *Tf-Idfweight* and *LSAweight* are the interpolation weights. These last two parameters were trained (see below) and used to control the impact of the *tf-idf* and *LSA* measures on the final ranking.

[9] v. 0.8.6 (http://infomap-nlp.sourceforge.net).

4.2.4. Parameter training

For training the parameters we have run the system by using 100 word descriptions taken from Burke *et al*. (1991). Every item in the training set consists of a target word and its corresponding description. For example, the word '*palindrome'* is represented in the training set by the following definition: *"word or sentence that reads the same backward or forward"*.

During training session we varied the value of three parameters: (a) number of the processed snippets (0 to 30), (b) *LSAweight*; and (c) *TfIdfweight.*

WF processed all descriptions of the training set, applying systematically all possible parameter combinations. The first hundred candidate words returned by *WF* were considered as hits. The word's position is equated with its occurrence in the output. If the target word did not occur at all in the output, we assigned it the position 100; last but not least we computed an average position of all target words.

The parameter combination yielding the *lowest average position* (*lap*) was considered to be the best combination. The following example might illustrate position assignment. Suppose the query and target word were respectively *"word or sentence that reads the same backward or forward"* and '*palindrome'*. If the output list contains the following set of ranked words, '*phrase*', '*spelled*', '*palindrome*', '*winston*', etc., then '*palindrom*e' is assigned the value 3, as it occurs in the third position.

In the training phase the following results have been achieved. Concerning the parameters, the best combination was 30 snippets (the largest set of possible queries via API authorized by Google), *LSAweight*: 0.5; *Tf-Idfweight*: 0.5; *lap*: 25.09. Table 1 summarizes the distribution of the positions of the target words processed under optimal parameter setting. The top row shows the word's position in the output list, while the bottom row shows the number of target words occurring in this position.

Position	1	2	3	4	5-7	8-15	16-45	46-99	100
# words	56	10	2	2	2	1	3	3	21

Table 1. Position distribution of target words in training set, using optimal parameters

Table 1 shows that 56 words occur in the first place, while 73 target words from the training set occurred in reasonable positions (<=15); 27 words occurred however very late.

Considering the best parameter combination separately for each description of the training set we can get an average position of 23.35, which is better than the result achieved with a unique set of parameters applied to all descriptions (25.09). Hence parameter fine-tuning for queries may be helpful. The *LSA* measure proved to be beneficial for ranking only 8 of the 79 target words found in the corresponding output lists, representing hardly more than 10%. Our explanation for this poor result is that

the target words from the training set were mostly low-frequency words, occurring too rarely in the corpus used for training the LSA matrix.

4.2.5. Evaluation

In order to empirically evaluate the system we first constructed a description set. First, we selected 100 "difficult" test words which have already been used by psychologists (Abrams *et al*. 2007) for TOT experiments. We asked human subjects (native and non-native speakers of English) to write a description or definition of these words. The experiment was performed online. The resulting set of descriptions was then used in order to evaluate our system, using the *lap* measure, as described in the previous section. The parameters were set to the values which proved best during the training stage: average position for the descriptions provided respectively by *native* speakers and non-*native* speakers: (32.40 *vs*. 29.36); Baseline (plain frequency ranking): 69.34. The results are given in Table 2.

Position	1	2	3	4	5-7	8-15	16-45	46-99	100
# words	14	9	2	1	2	4	4	6	54

Table 2. Position distribution of target words in test set

The average positions achieved for descriptions provided both by native and by non-native speakers (32.40 and 29.36 respectively) greatly exceed the baseline (69.34), and they do not differ a lot from the training result (25.09). However, Table 2 shows that only 46 from the 100 target test words were discovered by our system, and only 32 of them occurred on reasonable positions, compared to the 73 words in the training stage.

It should be noted though, that the training set was constructed from the descriptions provided by scientists for research goals, whereas the test set contains subject descriptions which can suffer various shortcomings (lack of accuracy and correctness). Taking a closer look at the test descriptions for which our system failed to produce the corresponding target words we discovered that most of these descriptions appeared to contain at least one of the following flaws: (a) incomplete description [emu: Ostrich]; (b) incorrect description [*Agony:* A psychic state of being inactive, unable to act, passive]; (c) joke attempts [*Castanets:* clicky things that beautiful Spanish women use in their hands to make music and excite horrible Spanish men].

5. Discussion and future work

As Rapp (2002) has shown, vector-based methods are well suited to reflect paradigmatic associations (such as synonymy). This is a highly relevant feature, since paradigmatically related words are often present in the authors' mind while the intended term is not. However, it is also known that such approaches are particularly sensitive to the occurrence frequency of a word in the training corpus (*cf.* Bullinaria

and Levy 2007). This is a very important point, as word finding problems generally occur with low frequency terms. However, LSA is a computationally demanding approach; the size of the training corpus as well as of the initial co-occurrence matrix has clear limits (which we have reached on our machines). For this reason, simpler, but broad-coverage approaches, like those applied by Sitbon *et al*. (2008) could turn out to be more appropriate for the purpose of finding TOT words.

Apart from quantity (and its implied processing complexity), other features of the training corpus are substantial as well: quality, diversity and adequacy. While large-scale corpora do nowadays exist, there is often a trade-off between quality and diversity. Moreover, a corpus can only convey adequate semantic relations (and therefore candidate terms) if it is topically related to the query. But how can this be assured in advance? These problems are not trivial, some of their solutions can probably only be approached empirically, that is, by experimenting and fine-tuning.

This is precisely something that we plan to do in the future. In addition, we see two lines of improvement for our system. First, for the time being, *Word Finder* cannot treat compounds or named entities such as *John Lennon* or *fairy tale*. In order to enable finding of multi-word expressions, one needs to include frequently recurring *n*-grams to the *WF* dictionary. Second, one could implement and combine additional similarity or ranking measures such as for example those based on WordNet or direct association measures.

A reasonable answer to the question that we posed in the title would probably be "No, not as such." The problems of complexity and corpus design prevent approaches like LSA from solving the TOT problem on a large and general scale. However, we have seen that our approach, taking into account other information as well, already proves to be beneficial. We think that a future solution to this problem has to combine various sources of information in an intelligent manner. And certainly, vector-based approaches should represent one of these sources.

References

ABRAMS, L., TRUNK, D.L. and MARGOLIN, S.J. (2007). Resolving tip-of-the-tongue states in young and older adults: The role of phonology. In L.O. Randal (ed.). *Aging and the elderly: psychology, sociology, and health*. Hauppauge, NY: Nova Science Publishers, Inc.: 1-41.

BADECKER, W., MIOZZO, M., and ZANUTTINI, R. (1995). *The two-stage model of lexical retrieval: Evidence from a case of anomia with selective preservation of grammatical gender*. Cognition, 57: 193-216.

BÉROULE, D. and ZOCK, M. (2009). Modelling the tip-of-the-tongue state in guided propagation networks. In B. Sharp and M. Zock (eds). *6th International Workshop on Natural Language Processing and Cognitive Science*. Milano: 77-93.

BOCK, K. and LEVELT, W. (1994). Language Production: Grammatical Encoding. Gernsbacher, M. (ed.) *Handbook of Psycholinguistics*. Orlando, FL, Academic Press: 945-984.

BOISSIÈRE, P. (1862). *Dictionnaire analogique de la langue française : répertoire complet des mots par les idées et des idées par les mots*. Paris: Auguste Boyer.

BROWN, A.S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin*, 109: 204-233.

BROWN, R. and MC NEILL, D. (1966). The tip of the tongue phenomenon. *Journal of Verbal Learning and Verbal Behavior*, 5: 325-337.

BULLINARIA, J.A. and LEVY, J.P. (2007). Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study. *Behavior Research Methods*, 39: 510-526.

BURKE, D., MACKAY, D., WORTHLEY, J. and WADE, E. (1991). On the tip of the Tongue: What Causes Word Finding Failures in Young and Older Adults? *Journal of Memory and Language*, 30: 542-579.

BUTTERWORTH, B. (1982). Speech errors: Old data in search of new theories. In A. Cutler (ed.). *Slips of the tongue and language production*. Amsterdam: Mouton: 73-108.

CARAMAZZA, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, 14: 177-208.

CEDERBERG, S. and WIDDOWS, D. (2003). Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003* – vol. 4, Edmonton, Canada: 111-118.

DEERWESTER, S.C., DUMAIS, S.T., LANDAUER, T.K., FURNAS, G.W. and HARSHMAN, R.A. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41 (6): 391-407.

DELL, G.S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, 93: 283-321.

DELL G., CHANG, F. and GRIFFIN, Z. (1999). Connectionist Models of Language Production: Lexical Access and Grammatical Encoding, *Cognitive Science*, 23/4: 517-542.

DONG, Z. and DONG, Q. (2006). *HOWNET and the computation of meaning*. World Scientific Publishing Co, Inc. River Edge, NJ., USA.

DUMAIS, S. (1990). *Enhancing Performance in Latent Semantic Indexing*. Technical Report TM-ARH-017527, Bellcore.

FAY, D. and CUTLER, A. (1977). Malapropisms and the structure of the mental lexicon. *Linguistic Inquiry*, 8: 505-520.

FELLBAUM, C. (ed.) (1998). *WordNet: An Electronic Lexical Database and some of its Applications*. Cambridge, MA: MIT Press.

FONTENELLE, Th. (1997). Using a bilingual dictionary to create semantic networks. *International Journal of Lexicography*, 10(4): 275-303.

FROMKIN V. (ed.) (1973). *Speech errors as linguistic evidence*. The Hague: Mouton Publishers.

JAMES, W. (1950/1890). *The principles of psychology*. Vol. One. New York, NY: Dover Publications, Inc.

KEMPEN, G. and HUIJBERS, P. (1983). The lexicalization process in sentence production and naming: Indirect election of words. *Cognition*, 14: 185-209.

KILGARRIFF, A., RYCHLY, R., SMRZ, P. and TUGWELL, D. (2004). The Sketch Engine. In: Williams,. G. and S. Vessier (eds). In *Proceedings of the 11th EURALEX International Congress*. Lorient: 105-116.

LANDAUER, T.K. und DUMAIS, S.T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review* 104: 211-240.

LANDAUER, T.K., FOLTZ, P.W. and LAHAM, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, 25: 259-284.

LEVELT, W., ROELOFS, A. and MEYER, A.S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22: 1-75.

LUND, K. and BURGESS, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behaviour Research Methods, Instruments and Computers* 28(2): 159-165.

MOERDIJK, F., TIBERIUS C. and NIESTADT, J. (2008). Accessing the ANW Dictionary. In Zock, M. and C. Huang (eds). *COGALEX workshop*, *COLING*, Manchester, UK.

RAPP, R. (2002). The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In *Proceedings of COLING*. Taipei, Taiwan.

REASON, J. and LUCAS. D. (1984). Using cognitive diaries to investigate naturally occurring memory blocks. In J.E. Harris and P.E. Morris (eds). *Everyday memory: Actions and absentmindedness*. London: Academic Press: 53-69.

RICHARDSON, S., DOLAN, W. and VANDERWENDE, L. (1998). Mindnet: Acquiring and structuring semantic information from text. In *ACL-COLING'98*. Morgan Kaufmann Publishers: Montréal: 1098-1102.

ROBERT, P., REY A. and REY-DEBOVE, J. (1993). *Dictionnaire alphabétique et analogique de la Langue Française*. Le Robert, Paris.

ROGET, P. (1852). *Thesaurus of English Words and Phrases*. Longman, London.

RUNDELL, M and FOX, G. (eds) (2002). *Macmillan English Dictionary for Advanced Learners* (MEDAL). Oxford.

SCHVANEVELDT, R. (ed.) (1989). *Pathfinder Associative Networks: studies in knowledge organization*. Ablex: Norwood, N.J.

SIERRA, G. (2000). The onomasiological dictionary: a gap in lexicography. In U. Heid, S. Evert, E. Lehmann and C. Rohrer (eds). *Proceedings of the Ninth EURALEX International Congress*. IMS, Universität Stuttgart: 223-235.

SITBON, L., BELLOT, P. and BLACHE, P. (2008). Evaluation of lexical resources and semantic networks on a corpus of mental associations. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis and D. Tapias (eds). *Proceedings of LREC'08*, Marrakech.

SMITH, S.M. (1994). Frustrated feelings of imminence: On the tip of the tongue. In J. Metcalfe and A. Shimamura (eds). *Metacognition: Knowing about knowing*. Cambridge, MA: MIT Press: 27-45.

SUMMERS, D. (1993). *Language Activator: the world's first production dictionary*. Longman, London.

T'ONG, T-K. (1862). *Ying ü tsap ts'ün* (The Chinese and English Instructor). Canton.

VIGLIOCCO, G., ANTONINI, T. and GARRETT, M. (1997). Grammatical gender is on the tip of Italian tongues. *Psychological Science*, 8: 314-317.

WADE-STEIN, D. and KINTSCH, E. (2003). *Summary Street: Interactive Computer Support for Writing*. Technical report, University of Colorado.

ZOCK, M. and SCHWAB, D. (2008). Lexical access based on underspecified input. In M. Zock and C. Huang (eds). *COGALEX workshop*, *COLING*, Manchester: 9-17.

Posters
and software demonstrations

DiCE in the web
An online Spanish collocation dictionary

Margarita Alonso Ramos[1], Alfonso Nishikawa, Orsolya Vincze
University of A Coruña

Abstract

DiCE is an online dictionary of Spanish collocations which provides semantic and combinatorial information of lexical units. The dictionary makes use of the typology of lexical functions (Mel'čuk *et al.* 1995), together with natural language glosses to describe the semantic content of collocates. With the aim of showing the different ways in which the database can be exploited, we present the organization of the online interface of the dictionary.

Keywords: collocation, online dictionary, lexical functions, Spanish as a foreign language.

1. Introduction

In this paper we present the *Diccionario de Colocaciones del Español* (DiCE), a web-based collocation dictionary of Spanish that is being developed at the University of A Coruña (Alonso Ramos 2005). Collocations in DiCE are idiosyncratic combinations of two lexical units, the *base* and the *collocate*, as defined by Hausmann (1979) and others. DiCE is similar to dictionaries such as the *BBI* (Benson *et al.* 1986), the *LTP* (Hill and Lewis 1997) or the *Oxford Collocations Dictionary* (Crowther *et al.* 2002). However, unlike these English paper dictionaries, it has been conceived from the start as an electronic lexical database. This allows us to provide more information to the user and to implement a flexible means of access to this information.[2]

As far as its theoretical framework is concerned, DiCE draws upon the fine-grained typology of lexical functions (LFs) introduced in the *Explanatory Combinatorial Lexicology* (Mel'čuk *et al.* 1995). However, users of the dictionary do no have to be familiar with this framework since the semantic content of LFs is paraphrased in natural language glosses.

2. The architecture of the dictionary

DiCE has been conceived as an electronic lexical database, a feature that makes it free from the alphabetical order of conventional dictionaries, given that the architecture of an electronic dictionary is necessarily a network, not a list. Our environment is divided

[1] lxalonso@udc.es.

[2] DiCE is maintained in a MySQL database and is implemented in PHP using an Apache Server and the CakePHP environment.

into two zones: the administration zone and the public zone. The first one is handled by the lexicographer; it is dedicated to the edition of lexicographic information contained in the microstructure and the macrostructure of the dictionary (*e.g.* semantic tags or the list of LFs). The public zone can be accessed freely by users. It consists of two main components: the dictionary itself and the advanced search component.

2.1. The dictionary component

We access the dictionary component through the list of lemmas. Each lemma is associated with a list of lexical units (LUs). For each LU, the user can look up the corresponding semantic or combinatorial information. As for the semantic information, the entry of each LU provides: a) a semantic tag that represents the generic meaning; b) the actantial structure representing the participants of the situation designated by the noun; c) corpus examples, most often derived from the online *Corpus of the Real Academia Española* (CREA); and d) quasi-synonyms and quasi-antonyms of the LU.

As for combinatorial information, we offer two sources of information: 1) the syntactic combinatory information of the LU is shown in the Government Pattern (*esquema de régimen)* section, where we specify the projection of its semantic valency structure onto its syntactic valency structure and, in addition, the subcategorization information associated with the latter, and 2) the lexical combinatory information is displayed in the section Collocations. In what follows, we focus on lexical combinatorics.

Taking a specific LU as the starting point, the user can choose between five different groups of lexical correlates:

1) Attributes of the participants: Under this heading, we have grouped those attributes or nouns that refer to the participants of the situation designated by the LU. For example, in the entry for ADMIRACIÓN ‘admiration’, the user finds *digno de admiración* ‘worthy of admiration’ or *admirable* ‘admirable’, both referring to the participant that can compel admiration;

2) LU + adjective. Here, the user finds adjectives that co-occur with the LU;

3) Verb + LU: In this section, we have grouped the verbs that take the LU as a direct complement or as a prepositional complement, *e.g. despertar antipatía* ‘[to] arouse dislike’;

4) LU + verb: This section contains verbs that take the LU as the grammatical subject, *e.g. el enfado se le pasó* ‘his anger subsided’;

5) Noun *de* LU: Here, we find noun collocates that precede the LU introduced by the preposition *de* ‘of’; *e.g. atisbo de esperanza* ‘a glimmer of hope’.

Once the user has entered one of these sections, he will find a list of collocates or semantic derivates preceded by an LF, a gloss, and followed by one or more examples. In the gloss we intend to give a brief indication of the meaning of the collocate in relation to the base. So, the gloss *intensa* ‘intense’ serves to group various adjectives

such as *ferviente* 'burning', *profunda* 'profound', and *enorme* 'enormous', which, in combination with the noun ADMIRACIÓN 'admiration', fulfill the same role, although they do not have strictly the same meaning. This proved to be a very useful feature especially for learners, who may have a problem choosing correctly between collocations which at first sight might appear to have similar meanings. For instance, the following adjectives used with the noun *admiración* are described in the glosses as follows:

(1) *incondicional*, glossed as *intensa* 'intense'

(2) *ciega*, glossed as *más intensa de lo conveniente* 'more intense than convenient'

(3) *general*, glossed as *compartida por muchos* 'shared by many persons'

(4) *eterna*, glossed as *que dura mucho* 'long-lasting admiration'

2.2. The advanced search component

The "Consultas avanzadas" ('advanced search') component serves principally to carry out specific searches. Rather than making queries for the collocates of a specific LU, it helps us find the answer for particular questions.

We can conduct three types of searches: 1) direct search, 2) inverse search and 3) writing aid.

2.2.1. Direct search

"Consultas directas" ('direct search') allows us to find the collocates of a base described by a given LF. Besides the LF, the user has a further option of specifying the lemma of the base and its lexical unit when carrying out a search (see Figure 1 for an example of a search for the collocates described by the LF Magn of the LU *estima 1b*).

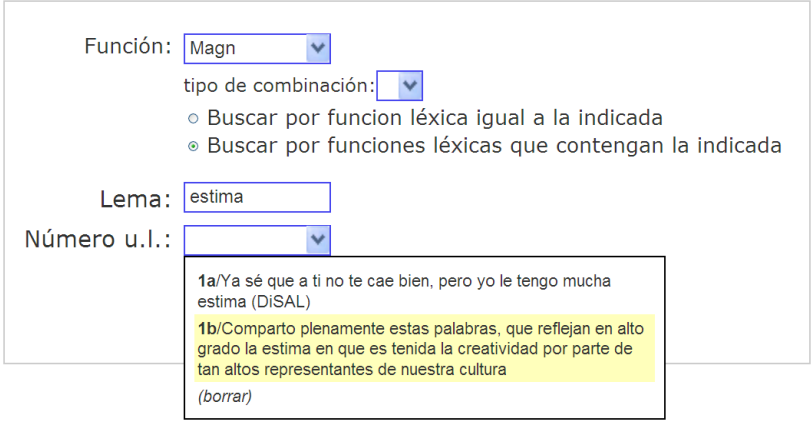


*Figure 1. Direct search for Magn(*estima *1b)*

2.2.2. Inverse search

We can conduct two types of searches using the option "Consultas inversas" ('inverse search'):

1) The first type of search allows us to find the base of a collocation starting from the collocate. After having indicated the collocate, we also have the option of specifying the LF associated with it. Figure 2 shows the results obtained from the search for the collocate *a raudales* 'in abundance'.

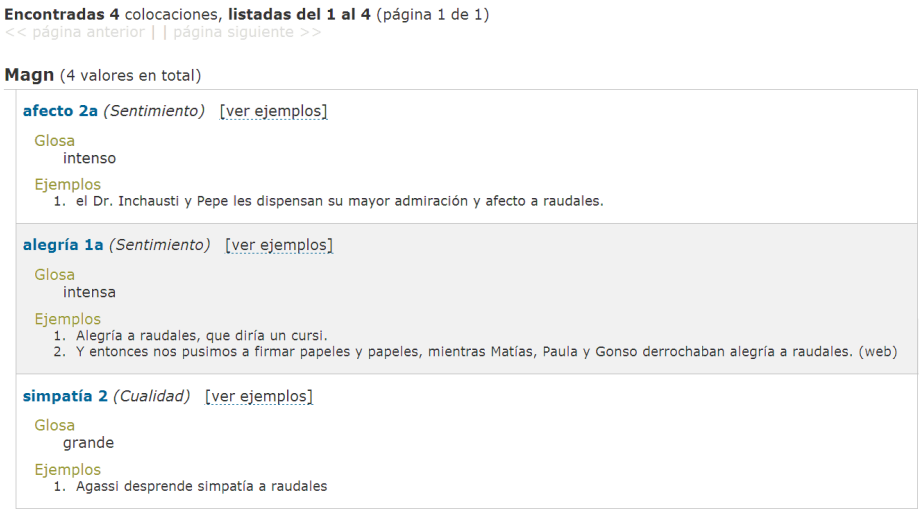


Figure 2. Results of an inverse search for a raudales *as a collocate*

2) The second type of search is more oriented towards comprehension. Here we can find out which LF – and gloss – codifies the relation between a given base and a collocate. For example, we can find that *a raudales* adds the meaning 'intense' to the base *alegría*.

2.2.3. Writing aid

The option "Ayuda a la redacción" ('writing aid') is intended to resolve questions concerning lexical combinatorics raised by any speaker of Spanish, including learners and native speakers. It helps us verify whether a certain combination of words is correct. At this moment, we offer the following two types of aid:

1) The first kind of aid allows the user to check whether a given base can co-occur with a given collocate (*cf.* Figure 3).

2) The second aid provides as search results collocates corresponding to a meaning, codified by a gloss, and a syntactic scheme (under "tipo"). Figure 4 shows a search for collocate adjectives of *alegría* 'joy', meaning 'caused by the misfortune of another person'.

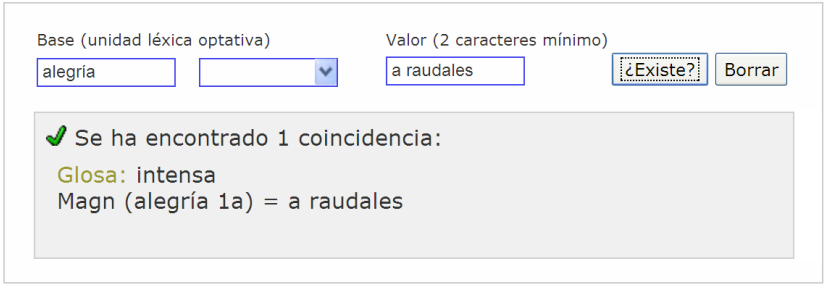


Figure 3. Checking collocations with the Writing aid tool

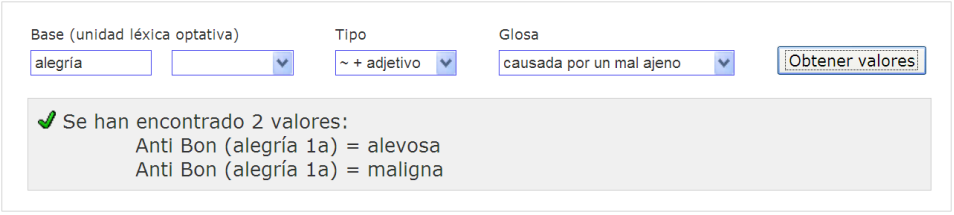


Figure 4. Finding collocates with the Writing aid tool

3. Conclusion and future work

As we have shown, the electronic format of DiCE and the codification of collocations through LFs and glosses turn out to be a clear advantage over conventional collocation dictionaries. The use of LFs allows for the efficient systematization of the representation of collocations. Such a systematic representation is an important aid for both the lexicographer and the users of the dictionary. DiCE, thus, provides structured information on Spanish collocations, which can be exploited in various ways by users through the different search options.

From a lexicographical point of view, along with a gradual expansion of the dictionary itself, one of our primary concerns is to look more closely into the possibilities of standardising the glosses and generalizing them in accordance with the meaning of bases.

In the mid-term future we also aim at exploiting the database integrating the dictionary with an exercise module, providing an online language learning environment. For further support of the learner, the next release of DiCE will furthermore offer each user the option to create his/her own learning space in which he/she can administrate personal collocation lists, annotations, performance scores and identified problems with respect to specific collocations or collocation types.[3]

[3] This paper was written within the framework of a research project: FFI2008-06479-C02-01 (Ministerio de Ciencia), and partially funded by FEDER.

References

ALONSO RAMOS, M. (2005). Semantic Description of Collocations in a Lexical Database. In F. Kiefer, G. Kiss and J. Pajzs (eds). *Papers in Computational Lexicography COMPLEX 2005*. Budapest: Linguistics Institute and Hungarian Academy of Sciences: 17-27.

BENSON, M., BENSON, E. and ILSON, R. (1998). *The BBI Combinatory Dictionary of English. A Guide to Word Combinations*. Amsterdam and Philadelphia: John Benjamins.

CROWTHER, J., DIGNEN, S. and LEA, D. (eds). (2002). *Oxford Collocations Dictionary for Students of English.* Oxford: Oxford University Press.

HAUSMANN, F.J. (1979). Un dictionnaire des collocations est-il possible? *Travaux de littérature et de linguistique de l'Université de Strasbourg*, 17/1: 187-195.

HILL, J. and LEWIS, M. (eds) (1997). *LTP Dictionary of Selected Collocations*. London: LTP.

MEL'ČUK, I., CLAS, A. and POLGUÈRE, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve: Duculot.

Tagging collocations for learners

Margarita Alonso Ramos[1a], Leo Wanner[2b], Nancy Vázquez Veiga[a], Orsolya Vincze[a], Estela Mosqueira Suárez[a], Sabela Prieto González[a]
[a]University of A Coruña, [b]ICREA and Pompeu Fabra University

Abstract

Collocations play a significant role in second language acquisition. In order to be able to offer efficient support to learners, an NLP-based CALL environment for learning collocations should be based on a representative collocation error annotated learner corpus. We are currently working on such a corpus for Spanish, starting from a fine-grained typology of collocation errors and drawing upon an existing learner corpus, namely CEDEL2 from the Autonomous University of Madrid. In this paper, we present this typology and discuss the first findings obtained from our annotation work.

Keywords: collocation, learner corpus, error typology, Spanish as second language.

1. Introduction

The importance of collocations in second language acquisition is increasingly recognized in the community (Lewis 2000; Granger 1998b; Howarth 1998; Nesselhauf 2003, 2005; Alonso Ramos 2006; Higueras 2006; Martelli 2006). To adequately support students in learning collocations, it is crucial to identify and classify the collocation errors made by them and then offer targeted exercises and adequate illustrative material. This presupposes the availability of collocation tagged learner and general corpora: a learner corpus allows us to identify the most common collocation errors; a general corpus is needed as a source of illustration and training material.

We aim at the development of an advanced NLP-based computer assisted language learning (CALL) environment for learning collocations in Spanish. In this paper, we focus on the problem of processing Spanish learner corpora, which consists of three stages: (i) analysis of the corpus and derivation of a collocation error typology; (ii) definition of a tag set to annotate the corpus; and (iii) tagging the corpus.

2. Towards a collocation error typology

A detailed analysis of learner corpora has proved to be essential (Dagneaux *et al*. 1998; Granger 1998a, 2007; Tono 2003). Such an analysis requires a predefined error

[1] lxalonso@udc.es

[2] ICREA and Pompeu Fabra University, leo.wanner@upf.edu. This paper was written within the framework of a research project: FFI2008-06479-C02-01 (Ministerio de Ciencia and partially FEDER).

tag set or error typology (Granger 2007). This is also true for the analysis of a collocation learner corpus. Currently available general learner error typologies tend to group collocation errors into a single subclass of lexical errors (Aldabe *et al*. 2005; Milićević and Hamel; 2007; Granger 2007; Díaz-Negrillo and García-Cumbreras 2007). Occasionally, collocation errors are also discussed referring to the POS of the collocation elements (Philip 2007). A closer look at a learner corpus reveals, however that a more detailed typology is needed. For the purpose of the present study, we used the *Corpus Escrito del Español L2* (CEDEL2) from the Autonomous University of Madrid[3], which consists of short compositions written by native speakers of English (L1). Consider some examples from CEDEL2:

(1) *deseo* lograr el gol *de ser bilingual*, lit. 'I desire achieve the goal of being bilingual'

(2) […] llenar un puesto [*de trabajo*], lit. 'fill a position [of work]'

(3) recibí un llamo *de Brad*, lit. 'I received a call from Brad'.

(4) *Algunos* tienen prejuicio *por edad*, lit. 'Some have prejudice for age'

Apart from errors not related to collocations (*e.g. bilingual* instead of *bilingüe*), which we ignore, the following collocation construction errors stand out[4]:

(1´) error in the base resulting from the projection of a word in L1 (English) to L2 (Spanish), *e.g. goal → gol*: *lograr* [*el*] *gol* – instead of *lograr* [*el*] *objetivo*;

(2´) error in the collocate resulting from a literal translation of a word from L1 to L2, *e.g. fill → llenar*: *llenar* [*un*] *puesto* – instead of *ocupar* [*un*] *puesto*;

(3´) error in the base resulting from a wrong morphological derivation and an inappropriate use of the collocation as a whole in the given context, *e.g. llamar → llamo*: *recibí un llamo de Brad* – instead of *recibí una llamada de Brad*; or, better: *me llamó Brad*;

(4´) error in the number of the base and in the governed preposition, *e.g. prejuicio*: *tienen prejuicio* [*por algo*], instead of *tienen prejuicios* [*hacia algo*].

[3] CEDEL2, which has been compiled by the group directed by Amaya Mendikoetxea, contains about 400,000 words of essays written in Spanish by native speakers of English. The essays are classified with respect to the proficiency level of the authors. The essays underlying our study were written by learners with intermediate or advanced level of Spanish. For more information, see http://www.uam.es/proyectosinv/woslac/cedel2.htm

[4] We interpret collocations in the sense of Hausmann (1979) as idiosyncratic word co-occurrences consisting of a base and a collocate.

The errors are very different. Therefore, a fine-grained collocation error typology is needed to capture these differences and be able to offer adequate didactic means to address them.

In the present stage of our work, we distinguish three main types of collocation errors: lexical errors, grammatical errors and register errors. Lexical errors concern either the whole collocation or one of its elements. In the first case, we find inexistent collocations in Spanish whose meaning would be correctly expressed by a single lexical unit (LU) (*e.g. *hacer de cotilleos*, lit '[to] make of gossip' instead of *cotillear* '[to] gossip'), and inexistent single LUs used instead of collocations (*e.g. *escaparatar* instead of *ir de escaparates*, lit. '[to] go of shop window'). In the second case, we distinguish between errors concerning paradigmatic lexical selection (*e.g. *lograr un gol* lit '[to] achieve a goal (in football)' instead of *lograr un objetivo,* lit '[to] achieve a goal') and errors concerning syntagmatic lexical selection (*e.g. *escribir el examen,* lit '[to] write the exam' instead of *hacer el examen,* lit '[to] do the exam); the former concern the base, the second the collocate.

Most lexical errors are literal translations from L1. Although a finer distinction is necessary later on to determine the source of errors, as a first approximation, the distinction between "transfer by importation", *i.e.*, adoption of an inexistent form in L2 – *recibir un llamo*, lit '[to] receive a call', instead of *recibir una llamada* – and "transfer by extension", *i.e.*, extension of the meaning of an L2 lexical unit – *salvar dinero*, lit. '[to] save money', instead of *ahorrar dinero* – is valid.

Grammatical errors in our typology are directly linked to collocations. They concern information that a learner cannot derive from the grammar of L2 and that must be described in the entry for the base of the collocation (*e.g. *hablar al teléfono* lit. '[to] speak to the phone' instead of *hablar por teléfono* '[to] speak through the phone').

In the class of register error, we group collocations that are pragmatically inappropriate. Thus, *tengo el deseo de ser bilingüe*, lit. 'I have the desire of being bilingual' sounds odd in an informal context – better: *me gustaría ser bilingüe* 'I would like to be bilingual'.

3. The process of tagging collocations in CEDEL2

Apart from a collocation error typology, a detailed semantic typology of collocations is crucial in order to be able to offer the learner examples of similar collocations. The most detailed and systematic semantically-oriented typology of collocations we know of are the Lexical Functions (Mel'čuk 1996), from now on referred to as LFs.

With the collocation error and the LF typologies at hand, we tag all collocations in CEDEL2. In the case of collocation errors, we also annotate the correct version of the erroneous collocation and the corresponding LF. Consider the following examples.

(1´´) *lograr* [*el*] *gol*: lexical error in the base; extension of the meaning of Sp. *gol* 'goal (in football)' due to phonetic similarity with Eng. *goal*; LF: Real1; correct: *lograr* [*el*] *objetivo*

(2´´) *llenar* [*un*] *puesto*: lexical error in the collocate; extension of the meaning of Sp. *llenar* 'fill' based on the English collocation [*to*] *fill a position*; LF: Oper1; correct: *ocupar* [un] puesto

(3´´) *recibí un llamo* [*de Brad*]: lexical error in the base; erroneous derivation based on the first person singular form of the verb Sp. *llamar*, possibly analogous with forms like *paseo<pasear*, *canto<cantar*, etc.; LF: Oper2; correct: *recibí una llamada* [*de Brad*]

Example (4'') shows that, on the one hand, a single collocation may show more than one error, and, on the other hand, that the determination of the source of an error is not always straightforward.

(4´´) *tienen prejuicio* [*por algo*]: 1. grammatical error in the government of the base; intralingual; 2. grammatical error in the number of the base; intralingual or possibly interlingual since Eng. *prejudice* can be used both as a countable or an uncountable noun; LF: Oper1; correct: *tienen prejuicios* [*hacia* algo].

The tagging of the learner corpus is currently being performed manually, supported by an interactive annotation tool, Knowtator, which is realized as a plug-in of the knowledge acquisition framework Protégé. The application allows us to define an annotation schema used in the process of annotation to give information on the semantics of the combinations – through LFs –, and, in the case of erroneous collocations, to describe the errors and propose a correction. We are also about to develop a collocation tagger that will tag both LFs and collocation errors. The work on the LF-tagger draws upon the work described in Wanner *et al*. (2006).

4. Conclusion

The preliminary evaluation of the corpus we annotated so far in accordance with the schema presented above reveals that 39% of the collocations used by learners contain some error. 62% of the erroneous collocations contain lexical errors, 33% show grammatical errors, whereas 5% have both lexical and grammatical errors. In a more fine-grained analysis of the more prominent lexical errors, we find that 54% of these represent an incorrect choice of the collocate, 20% the use of an incorrect base, 16% the use of an existing collocation with a different sense, while 10% are cases of using collocation-type constructions instead of single LUs. As for the possible source of errors, we can establish that the great majority of lexical errors – 70% – represent clear cases of lexical transfer from L1 to L2. However, further investigation based on a larger annotated corpus is needed to draw more fine-grained conclusions. We are thus currently working on the extension of our collocation error annotated learner corpus.

References

ALDABE, I., ARRIETA, B., DÍAZ DE ILARRAZA, A., MARITXALAR, M., ORONOZ, M. and URIA, L. (2005). Propuesta de una clasificación general y dinámica para la definición de errores. *Revista de Psicodidáctica*, 10/2: 47-60.

ALONSO RAMOS, M. (2006). Towards a dynamic way of learning collocations in a second language. In E. Corino, C. Marello and C. Onesti (eds). *Proceedings XII EURALEX International Congress*, Torino, Italy, September 6th-9th 2006. Alessandria: Edizioni Dell'Orso: 909-921.

DAGNEAUX, E., DENNESS, S. and GRANGER, S. (1998). Computer-aided error analysis. *System*, 26: 163-174.

DÍAZ-NEGRILLO, A. and GARCÍA-CUMBRERAS, M.A. (2007). A tagging tool for error analysis on learner corpora. *ICAME Journal*, 31/1: 197-203.

GRANGER, S. (ed.) (1998a). *Learner English on Computer*. London and New York: Addison Wesley Longman.

GRANGER, S. (1998b). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A.P. Cowie (ed.). *Phraseology. Theory, Analysis, and Applications*. Oxford: Clarendon Press: 145-160.

GRANGER, S. (2007). Corpus d'apprenants, annotation d'erreurs et ALAO: une synergie prometteuse, *Cahiers de lexicologie*, 91/2: 465-480.

HAUSMANN, F.J. (1979). Un dictionnaire des collocations est-il possible? *Travaux de littérature et de linguistique de l'Université de Strasbourg*, 17/1: 187-195.

HIGUERAS, M. (2006). *Las colocaciones y su enseñanza en la clase de ELE*. Madrid: Arco Libros.

HOWARTH, P. (1998). The phraseology of learners' academic writing'. In A.P. Cowie (ed.). *Phraseology. Theory, Analysis, and Applications*. Oxford: Clarendon Press: 161-186.

LEWIS, M. (2000). *Teaching collocation. Further developments in the lexical approach.* London: Language Teaching Publications.

MARTELLI, A. (2006). A corpus-based description of English lexical collocations used by Italian advanced learners. In E. Corino, C. Marello and C. Onesti (eds). *Proceedings XII EURALEX International Congress*, Torino, Italy, September 6th-9th 2006. Alessandria: Edizioni Dell'Orso: 1005-1012.

MEL'ČUK, I. (1996). Lexical Functions: A tool for the Description of Lexical Relations in the Lexicon. In L. Wanner (ed.). *Lexical functions in lexicography and natural language processing*. Amsterdam and Philadelphia: John Benjamins: 37-102.

MILIĆEVIĆ, J. and HAMEL, M-J. (2007). Un dictionnaire de reformulation pour les apprenants du français langue seconde. In G. Chevalier, K. Gauvin and D. Merkle (eds). *Actes du 29e Colloque annuel de l'ALPA tenu a l'Université de Moncton*, Moncton, Canada, November 4th-5th 2005. *Revue de l'Université de Moncton*, n° hors série: 145-167.

PHILIP, G. (2007). Decomposition and delexicalisation in learners' collocational (mis)behaviour. In M. Davies, P. Rayson, S. Hunston and P. Danielsson (eds). *Online Proceedings of Corpus Linguistics 2007*, Birmingham, United Kingdom, July 27th-30th 2007. Birmingham: University of Birmingham: 1-11.

NESSELHAUF, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 24/2: 223-242.

NESSELHAUF, N. (2005). *Collocations in a learner corpus.* Amsterdam and Philadelphia: John Benjamins.

NESSELHAUF, N. and TSCHICHOLD, C. (2002). Collocations in CALL: An investigation of vocabulary-building software for EFL, *Computer Assisted Language Learning*, 15/3: 251-279.

TONO, Y. (2003). Learner corpora: Design, development and applications. In D. Archer *et al*. (eds). *Proceedings of the Corpus Linguistics 2003*, Lancaster, United Kingdom, March 28th-31th 2003. Lancaster: Lancaster University, University Centre for Computer Corpus Research on Language: 323-343.

WANNER, L., BOHNET, B. and GIERETH, M. (2006). Making sense of collocations. *Computer Speech & Language*, 20/4: 609-624.

WWWJDIC
A feature-rich WWW-based Japanese dictionary

James Breen
Monash University, Australia

Abstract

The WWWJDIC WWW-based Japanese dictionary service is described. The service integrates many different dictionary functions into a single package, and contains many features aimed to assist language learners and readers of Japanese texts.

Keywords: online dictionaries, hyperlink, customization, dictionary access, CALL.

1. Introduction

The WWWJDIC WWW-based Japanese dictionary (Breen 2003) is an evolving multi-feature dictionary service based on free and public dictionary files. It is widely used in Japanese-language education, and has a number of functions specifically to aid language learners. The main server is at Monash University in Australia, and there are five mirror sites in Europe (2), North America (2) and Japan. Usage is currently at several hundred thousand accesses per day.

2. Design Goals

The dictionary service has been developed to research the implementation aspects of "tomorrow's dictionary" (Atkins 2002) which proposed a wide range of configurable features and options, such as customized layout, extensive hyperlinking to other resources, and integration of the usually distinct Japanese character and word dictionaries. In this respect the service aims to go well beyond the scope of the common commercial dictionary services based on copies of published bilingual dictionaries (Kenkyusha 2009, NTT 2009, Yahoo 2009).

The dictionary files used by the server are:

1. the JMdict/EDICT Japanese-English dictionary (Breen 2004a), which has about 140,000 entries;
2. the ENAMDICT dictionary of named entities, which has over 700,000 entries;
3. the KANJIDIC kanji (Chinese character) dictionary (Breen 2004b), which has detailed information on over 12,000 characters;

4. a collection of glossary files in fields such as life sciences, law, engineering, Buddhism and business.

Entries in the dictionaries can be accessed either by the Japanese headwords (either the kanji form or the pronunciation in the *hiragana* or *katakana* scripts) or by words in the glosses (see Figure 1). The kanji dictionary can be accessed via a variety of methods including the traditional radical/stroke-count and four-corner techniques, the character pronunciations, the character meanings, and various dictionary indices. A multi-component index based on the visual elements in the characters is particularly effective and popular. An external handwriting interface can also be used. The dictionaries are integrated so that a user, having found a particular character, can display word entries containing that character, or having selected a word, can examine the details of the constituent characters.

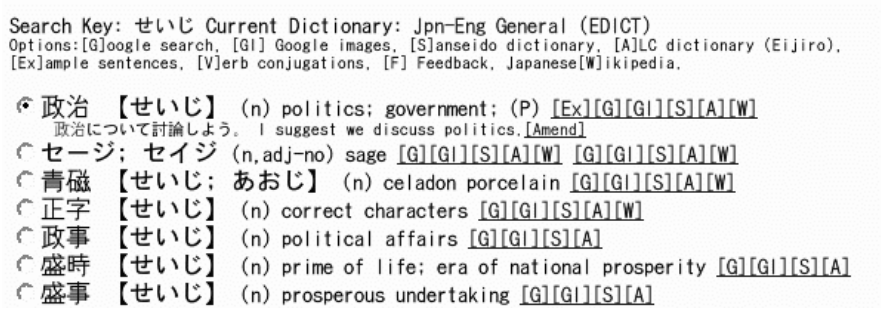


Figure 1. Example of dictionary word display

3. Text Glossing

One function of the service commonly used by students and translators is a text-glossing capability in which Japanese text is segmented and matched with dictionary entries. The segmentation and matching process uses a combination of most of the dictionary files, and allows inflected forms of verbs and adjectives to be aligned with the dictionary forms (*cf.* Figure 2).

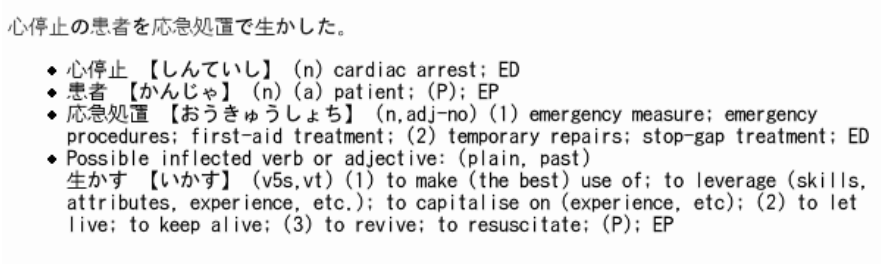


Figure 2. Example of text-glossing function

4. Features for Language Learners

Aspects of WWWJDIC's service which are of particular interest in CALL are:

- the option of displaying a table of conjugations for any of the verbs or verbal nouns in the dictionary (approximately 17,000 entries) (*cf.* Figure 3);
- animated stroke-order-diagrams for the 2,000 most common kanji;
- links at the entry level to the Tanaka Corpus of 150,000 Japanese-English sentence pairs (*cf.* Figure 4). The corpus can also be searched independently;
- sound clips of the Japanese pronunciation of almost all EDICT entries.

Conjugating 食べる: 一段 (type II) verb. (Supplementary Comments)

	Affirmative		Negative	
	Plain/Informal	Polite/Formal	Plain/Informal	Polite/Formal
Non-past	食べる	食べます	食べない 食べぬ(x) 食べず(に)(x)	食べません
Past	食べた	食べました	食べなかった	食べませんでした
Te-form	食べて	食べまして	食べなくて 食べないで	食べませんで
Conditional	食べたら	食べましたら	食べなかったら	食べませんでしたら
Provisional	食べれば	食べますなら(ば)	食べなければ	食べませんなら(ば)
Potential(#) & Passive	食べられる	食べられます	食べられない	食べられません
Causative	食べさせる 食べさす	食べさせます 食べさします	食べさせない 食べささない	食べさせません 食べさしません
Caus-Pass	食べさせられる	食べさせられます	食べさせられない	食べさせられません
Volitional/ Hortative	食べよう 食べる[よう/こと]にしよう	食べましょう 食べる[よう/こと]にしましょう	食べまい(+) 食べない[よう/こと]にしよう	食べますまい 食べない[よう/こと]にしましょう
Conjectural	食べるだろう	食べるでしょう	食べないだろう	食べないでしょう
Alternative	食べたり	食べましたり	食べなかったり	食べませんでしたり
Imperative	食べろ	食べなさい	食べるな	食べなさるな

Figure 3. Example of verb conjugation table

Other features of the service are:

- a configurable interface enabling users to structure the display and enable or disable options to suit their needs;
- multilingual operation. At present the main operating pages are available in English and Japanese. Other languages can be added by extending the catalogue files, and a French interface is in preparation;

- a restricted interface tailored for use with Japanese mobile phones;
- links from each entry to a range of online dictionaries, search engines, Japanese Wikipedia entries, the Japanese WordNet, etc;
- an edit interface enabling users to provide suggestions, amendments, etc. about dictionary entries or to propose new entries;
- an application program interface (API) enabling access from software and servers.

Although most of the dictionary files used are Japanese-English, it also includes the major WaDokuJT Japanese-German dictionary and smaller Japanese-French, Japanese-Spanish, Japanese-Swedish, Japanese-Hungarian, Japanese-Slovenian and Japanese-Dutch files.

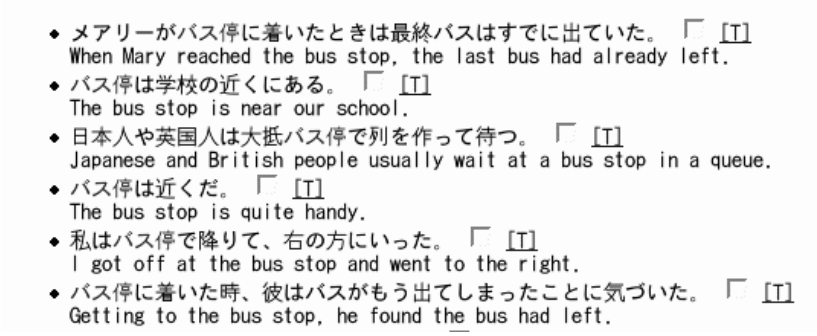


Figure 4. Example sentences linked to the バス停 entry

5. Conclusion

The WWWJDIC online Japanese dictionary server has demonstrated it is possible to develop an effective multi-faceted service which integrates many functions which are typically spread over several printed dictionaries, as well as providing features not available in traditional dictionaries. Future developments are planned to include a more dynamic “Web 2.0” user interface with greater user customization, a more interactive mode of operation, and dynamic information display.

References

ATKINS B.T.S. (2002). Bilingual Dictionaries – Past, Present and Future. In M.-H. Corréard (ed.). *Lexicography and Natural Language Processing – A Festschrift in Honour of B.T.S. Atkins*. Euralex 2002: 1-29.

BREEN J. (2003). A WWW Japanese Dictionary. In A. Tokita (ed.). *Language Teaching at the Crossroads*, JSC Working paper No. 13, Monash Asia Institute, Monash University Press, 2003. http://www.wwwjdic.net/.

BREEN J. (2004a). JMdict: a Japanese-Multilingual Dictionary. In G. Sérasset, S. Armstrong, C. Boitet, A. Popescu-Belis and D. Tufis (eds). *COLING Multilingual Linguistic Resources Workshop*, Geneva, 2004, 71-78.

BREEN J (2004b). Multiple Indexing in an Electronic Kanji Dictionary. In M. Zock and P. Saint Dizier (eds). *COLING Enhancing and Using Electronic Dictionaries Workshop*, Geneva, 2004, 1-7.

Kenkyusha LTD. (2009). *Kenkyusha Online Dictionary.* http://kod.kenkyusha.co.jp/service/.

NTT Resonant Inc. (2009). *Goo Jisho* http://dictionary.goo.ne.jp/.

Yahoo Japan Corporation. (2009). *Yahoo Jisho* http://dic.yahoo.co.jp/.

Have I got the wrong definition of…?

How to write simple technical definitions on the basis of examples taken from Newsgroup discussions[1]

Elisa Corino[2], Cristina Onesti[2]
Università di Torino

Abstract

This contribution is meant to point out the deficiencies of monolingual dictionaries concerning common technical language. Our study is based on the observation of some issues discussed in newsgroup messages dealing with Motor vehicles, in both Italian and English, which were taken from a subset of the NUNC (Newsgroup UseNet Corpora) suite of multilingual corpora. The study has revealed that technical meanings are often neglected in dictionary entries and users often run into difficulties. Such problems have a twofold explanation: on the one hand, entries are not really comprehensive of all the possible meanings; on the other hand, explanations are sometimes not clear enough. We investigate how definitions can be improved, pointing out some problems detected in dictionaries and reported directly by users in their posts, outlining a sample definition for technical terms by considering some of the key problems of lay users facing Language for Specific Purposes (LSP), and the strategies used by "experts" to make a term clear, as regards both structure and words used. The findings shed light on some problems of technical definitions in monolingual dictionaries and provide useful information to improve them.

Keywords: technical definition, LSP, newsgroups, bottom-up definitions.

1. Using Newsgroups in the dictionary making process

In a technoscientific text, terminology plays an essential role, being the core of conciseness, precision and appropriateness. It is common knowledge that technical terms often come from an ordinary lexicon which is turned into more specific subsenses in specialized fields (*e.g.* depression in economy or meteorology). A large number of studies points out the problems of those terminological fields which take part of their terminology from general language and give to these terms a specific entity. Ruiz Quemon (2006) for instance explored neuroscientific protocols observing

[1] We are grateful for the support of the VALERE Project (*Varietà Alte di Lingue Europee in Rete* – Formal Varieties in Newsgroups of European Languages: Structural Features, Interlinguistic Comparison and Teaching Applications), promoted by Regione Piemonte – Bando Scienze umane e sociali, responsible: Massimo Cerruti. The present paper is the result of a joint work by the two authors. However Elisa Corino is respondible for paragraph 2, Cristina Onesti is responsible for paragraphs 1 and 3.

[2] {elisa.corino,cristina.onesti}@unito.it

how common terms, which should be directly accessible to the users' experience, are often obscure and even ambiguous when used with a technically specialized meaning.

Recent online dictionaries have proved that new types of references evolving upward from readers directly onto the Net can be successful (see Ding 2008 and his "bottom up approach"). The ability of modern information retrieval systems to use dialogues with the user as feedback for improvements in dictionary making has been exploited by many web dictionaries. Their forums even managed to create online communities which share an immediate interest in the dictionary itself, suggesting misprints and typographical errors, problems with examples, erroneous or insufficient encyclopaedic information, missing entries, etc.

On the other hand, corpora and electronic resources have now entered the common praxis of dictionary making, helping lexicographers with frequency lists, providing them with examples in context and new entries. Nonetheless there is still little evidence of any approach combining both the "folk construction" of the dictionary and the exploitation of electronic resources.

This contribution is meant to give an example of how Newsgroup discussions collected in an annotated corpus can contribute to the dictionary making process, shading light onto the users' problems with Language for Specific Purposes (LSP) and, what is more, with the texts explaining the entries. This amounts to a "bottom up approach", where participants themselves ask for explanations and give detailed, though simple and understandable, definitions. The corpus is thus used as a source for entries and examples, as a sample for definitions and as a control tool. Newsgroup messages provide a sort of "natural definition" proposed by experts in such fields, who try to write in as clear a language as possible, although they still use a very specific terminology.

The NUNC (Newsgroups UseNet Corpora) is a suite of multilingual corpora (Italian, English, French, German, Spanish) developed at the University of Turin and made up of two years of the UseNet hierarchies. It is subdivided in different parts: a generic all-comprehensive corpus and three subcorpora related to Cooking, Motors and Digital Photography and thus presenting a high level of specificity. Experts in the three fields discuss technical matters with newcomers, showing a peculiar register variation (from formal to informal, see also Corino 2007). In this paper we analyze in particular the data extracted from NUNC-EN/IT-Motors.

2. *What's the definition of…?*: definitions in specialist newsgroups

As is well-known, in a specialist context, terminology comprehension, definition and translation are a thorny matter. If we focus on examples [1] and [2], we can easily notice how people need the support of "experts" to understand the subsenses of technical terms.[3]

[3] Please note that the examples have been accurately reproduced according to the original format.

[1] A: Or have I got the wrong definition of steer into ?

B: quite possibly. For example , if you're going around a right hand bend and the rear of the car breaks away towards the outside of the bend and starts to skid, steering into would be to steer left . i.e the car is breaking away to the left and you steer into the left . [NUNC-EN Motors]

A: Ahhh ... quite possibly indeed. I Am very glad I didn't get a chance to foul this up ... I had got the wrong definition !!

In [1] user A is quite disoriented by the definition of steer into he found so far and wonders whether it is correct. In fact, if we look up steer into in a common dictionary, what we find is an extremely broad sense which does not fit the users' expectations for a clarifying description:

verb 1 to control the direction in which a boat, car, etc. moves: [VN] He steered the boat into the harbour. (figurative) He took her arm and steered her towards the door 2 (of a boat, car, etc.) to move in a particular direction: [VN] The ship steered a course between the islands. [V] The ship steered into port. 3 [VN + adv./prep.] to take control of a situation and influence the way in which it develops (Oxford Advanced Learner's Dictionary 2000)

The answer given by B, by contrast, is very precise, covering the lack of the dictionary and filling A's need for a clear explanation. The feedback given by the user confirms the success of the definition and once more states the need for considering this meaning as well. It also highlights the need for an in-depth study of how the definition should be given.

Another example is the technical meaning of to bridge (an amp). Most monolingual (*e.g.* Oxford Advanced Learner's Dictionary 2000) and bilingual dictionaries (*e.g.* wordferefence.com) seem to ignore it and the user is forced to ask for some information:

[2] A: What does "bridging an amp" mean ? [MHa]

B: "Bridging" refers to taking two channels of an amplifier and combining them to turn the amplifier into a one channel amplifier. […] It should be clear that when you bridge an amp, you are changing *the amp*. The speaker's impedance is *not* a function of the amp, but the amp's tolerance to a given impedance depends completely on the way the amp is configured.

B gives a "dictionary-like" definition, then gives some more details which could be useful to his interlocutor. Here, as in [1], the presence of metadiscourse markers is not unimportant ("For example, if you're...", "*i.e.*", "It should be clear that"): they help the reader to follow the writer's reasoning. Metadiscourse markers "guide the reader through the maze of the writer's units of thoughts by indicating the organization of the text. On the other level, metadiscourse markers build an interaction between the reader and the writer and account for the atmosphere and reader-friendliness of the text" (Jalilifar and Alipour 2007); see also Crawford Camiciottoli (2003), Hyland (2005) and Parvaresh and Nemati (2008).

Similar examples can be found in the Italian corpus (NUNC-IT-Motori) where most of the technical terms are of English origin though being commonly used in Italian as well.

[3] A: scusate la mia ignoranza , ma cosè il Cruise Control ?

B: Per definizione dall'inglese Cruise , è la crociera , cruiser è l' icrociatore o la nave da crociera, ma viene definita così , anche una velocità media che può tenere un oggetto in movimento ... scherzo , volevo complicare le cose , è l' acceleratore automatico , quel dispositivo che ti permette di impostare una velocita , e tenerla senza dover premere sull' acceleratore , di solito si disinserisce al tocco di qualsiasi pedale (nel suo caso credo 2).[4] [NUNC-IT Motori]

In example [3], the initial joke is a kind of mockery of a traditional dictionary explanation, which points out how useless such tautologic and etimologic definitions are often perceived by users in practice. B explains the mockery and expressely affirms that he wanted to make things more difficult, but then he uses very simple and clear words which clearly contrast with his dictionary-like definition.

Another interesting example derives again from Italian car-related terminology:

[4] Freno Motore

A: che cos'è il freno motore ? si manifesta solo sui 4t o anche sui 2t? cosa cambia dall' iniezione ai carburatori, e perchè ormai tutte le varie CBR, GSXR ecc li hanno abbandonati? Cos'è la "trazione"? azz... una serie di domande impegnative se non hai idea del funzionamento dei motori .

B: Ci provo, magari qualcuno potrebbe corregermi se scrivo cazzate a quest'ora. 1 Allora: Buona volontà e capacità di espressione 110 e lode . Conoscenza tecnica: 90 / 100 1 Freno motore: l'azione frenante che ha un motore a gas chiuso. Nel 2T è molto meno evidente in quanto in ogni ciclo vi è una fase di lavaggio (passaggio dei gas dal carter alla camera di combustione) e quindi depressioni di scarsa entità. Nel 4T invece abbiamo una fase in cui le valvole sono completamente chiuse, se non c'è miscela da incendiare come nel caso di gas chiuso, la scintilla scocca ma non c'è scoppio e quindi nessuna spinta sul pistone: in questo caso , la discesa del pistone è frenata dal vuoto che esso stesso crea nel cilindro andando verso il PMI e quindi il motore tende a frenare. 1 Vediamo un po' di precisare meglio il concetto. A prima vista sembra giusto, ma cozza contro un fenomeno fisico per cui è vero che il pistone mette in depressione il cilindro a valvole chiuse, ma è altrettanto vero che la stessa depressione richiama poi il pistone verso l'alto. Il vero effetto frenante è dato dallo stesso fenomeno, determinato più a monte e durante la fase precedente: l'aspirazione . […][5] [NUNC-IT Motori]

[4] Eng.: Sorry for my ignorance, but what is the Cruise Control? As for its definition it comes from the English Cruise, that is *la crociera*, cruiser means *incrociatore* or *nave da crociera* (cabin cruiser), but it is defined as follows: the average speed a moving object can have… I'm joking, just wanted to make things more complicated, it is the automatic accelerator, that A device that enables you to set the speed and keep it without having to push on the accelerator, it usually switches off when you touch another pedal.

[5] Eng.: What's the motor mounted brake? does it only work on the 4 stroke or also on the 2 stroke engines? What's the difference with the carburator injection and why have all various CBR, GSXR

Also in this case, monolingual dictionaries (see Dizionario interattivo Garzanti 2005) cannot satisfy the need for a precise terminological help nor provide contexts without ambiguities. User B mentions different cases and tries to summarize the content, rewording his definition and making it more precise (*Nel 4T invece, Vediamo un po'di precisare meglio, ma è altrettanto vero…*).

3. Conclusion

This contribution is a preliminary study meant to give some suggestions on how Newsgroup UseNet Corpora discussions and data can be used to shed light on users' expectations, to provide authentic and up-to-date materials following simple explanations and an understandable language despite LSP difficulties.

Newsgroups' data give some interesting hints to lexicographers, highlighting users' need for something more than a bare definition: they prefer to get details about the actual functioning of the object to be defined. Practical examples have proved to be more useful and in many cases addressed directly to the user (*e.g.* "For example, when you...").

It would be desirable to exploit such corpus data in the future within a bottom up approach, thus developing a tool which would be closer to single technical definitions.

References

CORINO, E. (2007). "NUNC est disputandum". In E. Barbera, E. Corino and C. Onesti (eds). *Corpora e linguistica in rete*. Perugia: Guerra: 225-252.

CRAWFORD CAMICIOTTOLI, B. (2003). Metadiscourse and ESP reading comprehension: An exploratory study. *Reading in a Foreign Language*, vol. 15, no 1: 28-44.

DING, J. (2008). Bottom-up Editing and More: The E-forum of The English-Chinese Dictionary. In E. Bernal and J. DeCesaris (eds). *Proceedings of the XIII Euralex International Congress* (Barcelona, July 15th-19th, 2008), IULA: 339-344.

HYLAND, K. (2005). *Metadiscourse: Exploring Interaction in Writing*. London, Continuum.

HORNBY, A.S. and WEHMEIER, S. (eds) (2000). *Oxford Advanced Learner's Dictionary*. Oxford: Oxford University Press.

---

etc. abandoned them? what is "traction"? damn... a lot of demanding questions if one has no idea of how engines work.

B: I'll try, someone might correct me if I write some shit at this time. 1 So: good will and expression 110 com laude. Technical competence: 90/100 1 motor mounted brake: the braking power of a closed gas motor. In the 2 stroke engine it is much less evident as in each cycle there is a washing phase (gas passage from the carter to the combustion chamber) and scarce depressions. In the 4 stroke engine we have a phase where all the valves are totally closed, if there is no mixture to burn as it happens with the closed gas, the spark shoots out but there is no burst and no push on the piston: in this case the piston descent is braked by the vacuum created by itself in the cylinder when going towards the PMI, therefore the engine tends to brake. 1 Let's try to specify the concept. At first sight it seems right, but it clashes against a physical phenomenon as it is true that with closed valves the piston depresses the cylinder, but it is also true that the same depression recalls the piston upwards. The real braking effect is due to the same phenomenon, determined during the aspiration phase.

JALILIFAR, A. and ALIPOUR, M. (2007). How explicit instruction makes a difference: metadiscourse markers and EFL learners' reading comprehension skill. *Journal of College Reading and Learning*.

PARVARESH, V. and NEMATI, M. (2008). Metadiscourse and Reading Comprehension: The Effects of Language and Proficiency. *Electronic Journal of Foreign Language Teaching*, vol. 5, no. 2: 220–239.

RUIZ QUEMOUN, F. (2006). L'hermetisme terminologique des protocoles en neurosciences. In E. Corino, C. Marello and C. Onesti (eds). *Proceedings of 12th EURALEX International Congress, Euralex 2006* (Torino, Italy, September 6-9, 2006). Alessandria: Edizioni Dell'Orso: 831–836.

Dizionario interattivo Garzanti 2005: www.garzantilinguistica.it

www.wordreference.com

Some editorial orientations for a multi-tier electronic monolingual school dictionary

Nathalie Gasiglia[1]
U.M.R. STL, Université Lille 3

Abstract

Presenting dictionary texts in electronic format offers important perspectives for enriching information and enhancing readability. The editorial orientation I propose consists of a tiered system of entries with simple texts for beginners and more substantial information for older pupils, with explicit semantico-syntactic patterns, and with an original way of presenting explicit mark-up, and synonyms and antonyms.

Keywords: electronic dictionaries, school dictionaries.

1. Introduction

As part of their school learning in France, pupils of 6 to 14 can use a wide range of printed monolingual dictionaries, but very few electronic ones: while twenty or so printed dictionaries for learners are published by Auzou, Hachette, Larousse, Le Robert, etc., only two electronic versions exist (designed for pupils of 8 to 10) published by Auzou and Le Robert. This situation may change now that primary and secondary schools are making increased use of electronic media. As part of a more general analysis of electronic dictionaries, this paper aims to present three editorial orientations designed to improve access to linguistic information at school.[2]

2. A multi-tier system

Two observations have led me to try and find out how an electronic dictionary might evolve as the skills of those consulting it develop:

(i) during their first years at school, pupils should be able to access an increasing amount of information formulated in terms adapted to their level of intellectual maturity, as is the case with the range of printed dictionaries available;

(ii) when working with pupils who are learning to read, teachers make considerable efforts to help them decode information in the dictionaries designed for them (this is reflected in schoolbooks), but this kind of assistance is given less and less frequently as dictionaries become more complex: it is thus hardly surprising that not all high school pupils know how to make the best use of their dictionaries.

[1] U.M.R. 8163 du C.N.R.S. (STL) & Université Lille 3 (France), nathalie.gasiglia@univ-lille3.fr
[2] See also Gasiglia (2009: 267-289) for other developments about these questions.

The editorial orientation I propose consists of a tiered system of entries with simple texts for beginners and more substantial information for older pupils, maintaining the visual pointers that help the lookup process (indicators, symbols, etc.). This would enable each user to consult a type of entry whose scope and degree of complexity is best adapted to his or her skills and to the type of information required.[3]

Version for the youngest users:

initier et **s'initier** verbe ☞ Toutes les formes du verbe

❶ Une personne ou un texte initient (une personne) à une activité [un travail, un art, un sport] quand ils lui en enseignent les connaissances élémentaires.
Le nouveau professeur de musique initie les élèves à la pratique du banjo.
Des exercices du manuel de français initient les élèves à la manipulation des dictionnaires.

↳ Une personne s'initie à une activité [un travail, un art, un sport] quand elle en apprend les connaissances élémentaires.
Léa s'initie-t-elle au piano seule ou avec un professeur ?
☞ Ce qu'il faut savoir pour exprimer l'idée de **l'initiation** d'une personne à quelque chose

❷ Une personne ou une chose initient une action ou un événement quand elles les déclenchent.
Le maire a initié la construction d'une nouvelle école.
☞ Ce qu'il faut savoir pour exprimer l'idée d'une **initiative prise** ou de la **cause** d'un événement

Version for high school pupils:

initier et **s'initier** verbe ☞ Toutes les formes du verbe

❶ Le verbe *initier* a été emprunté peu après 1350 au latin *initiare*, qui signifiait « enseigner les premiers éléments de (quelque chose) ». Le verbe français, comme son étymon latin, s'est d'abord appliqué aux mystères religieux avant de s'étendre à d'autres domaines.

◊ Descriptions triées par ordre historique

1 Une personne ou un texte initient (une personne) à une activité [un travail, un art, un sport] quand ils lui en enseignent les connaissances élémentaires.
Le nouveau professeur de musique initie les élèves à la pratique du banjo.
Des exercices du manuel de français initient les élèves à la manipulation des dictionnaires.

2 Une personne ou un texte initient (une personne) à un savoir ésotérique quand ils lui en transmettent les connaissances fondamentales.
Les alchimistes du Moyen Âge initiaient leurs disciples aux secrets de la matière.

3 Une personne ou un texte initient (une personne) à un culte quand ils lui transmettent des connaissances qui fonderont sa croyance.
Différents livres initient aux religions d'Extrême-Orient.

↳ Une personne s'initie à une activité [un travail, un art, un sport], à un savoir ésotérique ou à un culte quand elle en apprend les connaissances élémentaires.
Léa s'initie-t-elle au piano seule ou avec un professeur ?
☞ Ce qu'il faut savoir pour exprimer l'idée de **l'initiation** d'une personne à quelque chose

❷ Le français a étendu les emplois du verbe *initier* en empruntant le sens « être à l'origine de » du verbe anglais *initiate*.
Une personne ou une chose initient une action ou un événement quand elles les déclenchent.
Le ministre de l'environnement a initié un nouveau programme de recyclage des déchets.
☞ Ce qu'il faut savoir pour exprimer l'idée d'une **initiative prise** ou de la **cause** d'un événement

Figure 1. A tiered system of entries

[3] See also Atkins (2002: 12-13) and Verlinde *et al.* (2009).

For instance, *cf.* Figure 1:

1) the information displayed for 6-8 year olds would express definitions and contextualizations in simple terms, while for 8-10 year olds and younger high school pupils it would present further semantic splits and use more sophisticated vocabulary and syntax;

2) the version for the youngest users would limit the display of etymology to borrowings whose phonographic properties are not consistent with French norms, pointing out these unusual characteristics and briefly describing the foreign origin of the items in question; the version for older primary pupils would give elementary indications for all borrowed words (only adding etymology to information on pronunciation and spelling when appropriate); while the version for high school pupils would extend the information to words inherited from Latin or constructed in French (linking these local indications to the etymological, historical and morphological information given throughout);

3) the different meaning descriptions are given here in order of complexity (from most general to most specialised usage), this view can be updated on demand to reflect their historical order or to explicitate the derived meanings described in 3, 2 and 1 of sense 1 (*cf.* "Version for high school pupils").

3. Helping pupils identify semantico-syntactic patterns in context

Pupils often have an inadequate understanding of the way predicates (especially verbs) bring syntactical and semantic constraints to bear on their context. Designing clear presentational templates for constructional patterns involves creatively exploiting the potential of the electronic medium.

3.1. To make patterns easier to identify

To help pupils to understand the word *initier* in a sentence like *Le projet Bus Théâtre* ***est initié*** *par le club d'art dramatique de l'école*, the electronic dictionary could propose a number of questions in order to find the most appropriate description:

For high school pupils (cf. Figure 1):

Le verbe *initier* est-il employé à la voie passive (*être initié*) ?

↓ [si oui alors]

Le sujet du verbe évoque-t-il un humain ?

↓ [si non alors]

Voir description de sens n°2.

3.2. To make patterns easier to understand and re-use

When printed dictionaries present set patterns, space limitations mean that they can be both highly coded and incomplete; moreover, for verbs with complex patterns it is often hard to correlate the coded representation with the contextualisations given. To make patterns easier to understand and re-use, I feel it would be appropriate to link them ex-

plicitly to definitions and contextualisations, thus making it easier for users to identify precisely what corresponds to each element of the pattern. By way of example,

• for *initier* when it means *proposer une initiation* (*to initiate*), contextualisations would illustrate one pattern (*cf.* Figure 1):

subject = qqn **initie** direct object = qqn indirect object = à qqch.
Le nouveau professeur de musique **initie** les élèves à la pratique du banjo. (= Il les y initie)

• for *permettre* when it means *autoriser* (*authorize*), contextualisations would illustrate three patterns:

Qqn **permet** (qqch à qqn + à qqn de Vinf + qu P)
Le médecin **permet** le chocolat à Léa. (= Il le lui **permet**)
Son père **permet** à Luc de sortir. (= Il le lui **permet**)
Le maître **permet** que nous jouions. (= Il le **permet**)

This innovation, which could significantly reinforce the metalinguistic skills of pupils, nevertheless means taking into account the way similar information is presented in coursebooks: for this reason visual codes such as highlighting, underlining, frames and colours could be set to match what the pupils are used to seeing elsewhere.[4]

4. Helping pupils express things in different ways

To enable the dictionary to fully realise its potential as a production aid, an analysis of the possibilities offered by electronic media has inspired an original way of presenting explicit mark-up, and synonyms and antonyms.

1) Stipulating a variety of lexical usages makes it easy to know normative positions:

ⓘ L'emploi du verbe *initier* dans le sens "être à l'origine de" est relativement fréquent, mais il est déconseillé parce que c'est un emprunt à l'anglais qui n'est pas indispensable à l'expression en français. Il ne doit pas éclipser d'autres verbes de même sens ou qui apportent des nuances intéressantes dans certains contextes. (cf. Figure 1, meaning 2)

2) Explicitly linking synonyms and opposites to contextualisations immediately makes clear any adjustments required when substitution takes place. Each contextualization could have a pull-down menu presenting correlates that can directly replace the item in question.

[4] See also *Base lexicale du français* (http://ilt.kuleuven.be/blf/) and Verlinde *et al.* (2004: 428-432) for other presentations.

Le ministre de l'environnement	*a initié* ▼ *un programme de recyclage des déchets.*
	a amorcé
	a déclenché
	[…]

Further synonymic and antonymic reformulations could be presented as coherent subsets linked to explicit reformulation notes, according to whether they include:

– synonyms and opposites requiring syntactic remodelling:

• *Qqn* ***permet*** *à qqn de Vinf* / *Qqn* ***autorise*** *qqn à Vinf*

Son père ***permet*** *à Luc de sortir*. / *Son père* ***autorise*** *Luc à sortir*. *(= Il l'y* ***autorise****)*

• *Qqn* ***initie*** *qqch* / *Qqn* ***donne un coup d'arrêt*** *à qqch*

Le ministre de l'environnement ***a initié*** *un nouveau programme de recyclage des déchets.*

Le ministre de l'environnement ***a donné un coup d'arrêt*** *au nouveau programme de recyclage des déchets.*

– words that are morpho-semantically linked, such as nominal derivatives of verbs in support-verb constructions:

• *Qqn* ***initie*** *qqch* / *Qqn* ***prend l'initiative*** *de qqch*

Le ministre de l'environnement ***a initié*** *un nouveau programme de recyclage des déchets.*

Le ministre de l'environnement ***a pris l'initiative*** *d'un nouveau programme de recyclage des déchets.*

• *Qqn* ***permet*** *à qqn de Vinf* / *Qqn* ***donne*** *à qqn* ***la permission*** *de Vinf*

Son père ***permet*** *à Luc de sortir*. / *Son père* ***donne*** *à Luc* ***la permission*** *de sortir.*

– alternative syntactic constructions: *cf. permettre* in § 3.2.;

– alternative constructions (passive for example):

Un programme de recyclage des déchets ***a été initié*** *par le ministre de l'environnement.*

– alternative phrases:

Il ***est permis*** *à tout le monde de se tromper ! / Tout le monde peut se tromper ! / L'erreur est humaine ! / etc.*

5. Conclusion

Presenting dictionary texts in electronic format offers important perspectives for enriching information and enhancing readability. Though the editorial orientations pre-

sented here do not cover all possible fields of investigation, they touch on three areas where dictionaries might make significant advances by changing their lookup medium in order to help young learners absorb linguistic codes more effectively.

References

ATKINS, B.T.S. (2002). Bilingual dictionaries: past, present and future. In M.-H. Corréard (ed.). *Lexicography and Natural Language Processing. A festschrift in honour of B.T.S. Atkins*. EURALEX: 1-29. – 1e ed. (1996). In M. Gellerstam, J. Järborg, S.-G. Malmgren, K. Norén, L. Rogström and C. Röder Papmehl (eds). *Euralex'96 Proceedings*. Göteborg: Göteborg University: 515-546.

GASIGLIA, N. (2009). Évolutions informatiques en lexicographie : ce qui a changé et ce qui pourrait émerger. *Lexique*, 19: 235-298.

VERLINDE, S., SELVA, T., PETIT, G. and BINON, J. (2004). Les schémas actanciels dans le dictionnaire: point de convergence entre la morphologie et la sémantique lexicale. In G. Williams and S. Vessier (eds). *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6-10, 2004*. Lorient: Université de Bretagne-Sud: vol. 2, 427-436.

VERLINDE, S., SELVA, T. and BINON, J. (2009). Les bases de données au service d'un dictionnaire d'(auto-)apprentissage pour allophones. *Lexique*, 19: 217-233.

Automatic annotation of actants in specialized corpora

Fadila Hadouche[1], Marie-Claude L'Homme[2], Guy Lapalme[1]
Université de Montréal

Abstract

The objective of our work is to develop an automatic method for identifying actants (also called *arguments*) of predicative lexical units in running text. This work is carried out within a larger project that aims at providing rich contextual information in terminological databases. More specifically, the project consists in annotating predicative terms, *i.e*. verbs, and their participants in contexts extracted from a French corpus of texts on computing and the Internet. The participants are divided into two groups: actants that are required to define the lexical unit and circumstants that are optional.

Keywords: actancial annotation, semantic annotation, syntactic rules, specialized corpora, terminological database.

1. Introduction

It is increasingly recognized that reference works (general-language as well as specialized dictionaries) should include information about the syntactic structures in which lexical units or terms can be found. In order to do this, different models have been proposed in lexical databases (*e.g*. Wordnet provides what is referred to as sentence frames for specific lexical units; FrameNet, which describes lexical units within semantic frames, supplies a large number of annotated contexts, along with valency patterns). These ressources are useful for linguistic as well as Natural Language Processing (NLP) applications like question/answering, information extraction, translation, and information retrieval.

In this article, we present a project that aims at providing detailed information about the argument structure of predicative lexical units. More specifically, our aim is to annotate predicative terms along with their participants in contexts extracted from a French corpus of texts on computing and the internet. This annotation includes the following information: verbal lexical unit, semantic roles (*e.g*. agent, patient, destination, instrument, manner, location and means), syntactic functions (*e.g*. subject, object, complement), syntactic groups (*e.g*. noun phrase, adverbial phrase, prepositional phrase), and the indication of the type of participant (actant or circumstant).

[1] RALI, Université de Montréal
[2] OLST, Université de Montréal

Figure 1 shows how the lexical unit ACCÉDER is annotated together with the participants identified: [processeur] is labeled as an agent actant; [directement] is a circumstant of manner; and [cache] is labeled as an actant of location. Participants are divided into two groups: (1) actants that are necessary to define the lexical unit; and (2) circumstants which are optional (Mel'čuk 2004). As shown in Figure 1, the participants are in brackets. Their syntactic function, syntactic group and role are given in subscripts. Types of actant and circumstant are written respectively ACT and CIRC, syntactic group like SN for noun phrase, SAdv for adverbial phrase, etc. and syntactic function like SUJ for Subject, OBJ for Object, MOD for modifier, COMPL for complement are written in italics and semantic roles are written in upper case. The lexical unit is written in bold and in upper case.

Le [processus]$_{(SN,\ SUJ,\ \text{ACT, AGENT})}$ **ACCÈDE** [directement]$_{(SAdv,\ MOD\ \text{CIRC, MANIÈRE})}$ [au cache primaire]$_{(SP,\ COMPl\ \text{ACT,, LIEU})}$

In English: The process accesses directly to the primary cache

Figure 1. Example of manual annotation

It is worth mentioning that the methodology used to annotate the contexts manually is largely based on that developed within the FrameNet project (Ruppenhofer *et al.* 2006). However, in our work, while participants are labeled in terms of semantic roles, in FrameNet, they are labeled as Frame elements. This annotation is a time-consuming task that prompted us to develop an automatic process to annotate participants, more specifically actants. We are convinced that automating part of the process will also provide a valuable assistance to terminologists carrying out this task.

In this work the focus is on French verbal lexical units. Our corpus is composed of 105 lexical units and 2309 sentences manually annotated by terminologists. Our method aims (1) to identify participant actants by using rules extracted in the parser for French "Syntex"; and (2) to annotate the actants identified during the previous step with semantic roles. In this experiment, we annotate the roles of agent and patient.

2. Automated method for finding relevant participants in contexts

As was mentioned above, the first part of our method consists in locating relevant participants in contexts; then, we can attempt to label them. These tasks are further divided into three steps: identification of participants, type assignment of participants and semantic roles identification. These tasks are based partly on the extraction of rules using the syntactic links between the lexical units and their participants.

2.1. Rule extraction

During the first step, we use the Syntex parser (Bourigault *et al.* 2005), which computes the dependencies between the components of a sentence, as shown in Figure 2, in which the verbal lexical unit is in uppercase and syntactic group labels are in italics (Det stands for determinant, Nom for noun, VPpa for past participle verb,

Adv for adverb, Prep for preposition). Syntactic functions are in uppercase on the arrow links.

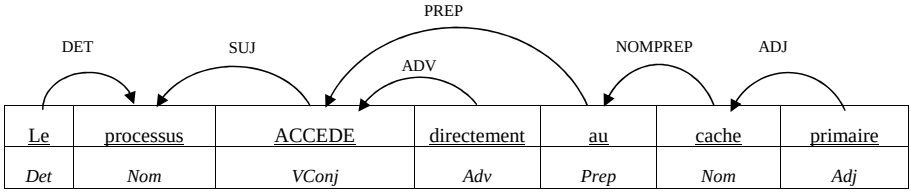


Figure 2. Syntex dependencies

In the results produced by Syntex, we focus on left and right dependencies of the verbal lexical unit ACCÉDER. From there, we extract rules to identify participants and their features. This is performed by combining the information provided by Syntex (Figure 2) with the manual annotation of participants (Figure 1). An example of the combination is shown in Figure 3. (On top of the sentence, we see the Syntex links of a sentence components indicated by arrows and below, we see the manually annotated participants in our corpora shown with dotted arrow.) This information will serve as a basis to build rules (*cf.* Section 2.1.1).

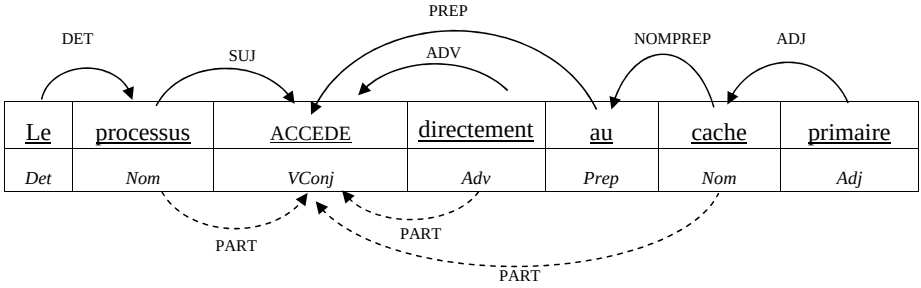


Figure 3. Combination schemata

2.1.1. Rules for participants identification

The left hand side of the rule (conditions of application) is composed of the words appearing in the sentence with the top links of Figure 3. The right hand side states the participants identified. In a rule, a word is described by its features <wordi, Posi, Rolei>, in which **Posi** is the part of syntactic group of **wordi** and **Rolei** is its syntactic function. The lexical unit is in uppercase and the result in the right hand is described by the features corresponding to the participant word. These rules are shown in (1), (2), (3) and (4):

Lexical unit + <$word_i$, *Nom, SUJ*> → <$word_i$,*participant,Sujet* > (1)

Lexical unit + <$word_i$, *Nom, OBJ*> → <$word_i$,*participant, Objet*> (2)

Lexical unit + <$word_i$, *Adv, ADV*> → <$word_i$,*participant, adverbe*> (3)

Lexical unit + <$word_i$,*Prep,PREP*> + <*word, Nom, NOMPREP*>

→ <$word_i$, *participant*,NOMPREP> (4)

2.1.2. Assigning type actant or circumstant

The feature assigns a type to the participant: actant or circumstant. The features <Nom, SUJ> and <Nom, OBJ> are actants in all cases (with all verbs) and the feature <Adverbe, ADV> is always associated with a circumstant (with all verbs). *i.e.* 'le processus ACCEDE directement' and 'le processus ABANDONNE directement'; [processus] in subject position is an actant for the two verbs ACCEDER and ABANDONNER. The adverbe [directement] is a circumstant for these two verbs. Here, the type actant and circumstant depends only on the features (corresponding to subject, object, or adverb) no matter what verbal lexical unit is used. In this case, the corresponding rules are shown in (5), (6), (7) and (8)

<*word,participant,SUJ*> → <*word, actant,SUJ,role*> (5)

<*word, participant,OBJ*> → <*word, actant,OBJ,role*> (6)

<*word, participant, ADV*> → <*word, circumstant,ADV, role*> (7)

<*word,participant, NOMPREP*> → <*word,type,role*> (8)

In (8), "type" can be actant or circumstant. In the example given in Figure 1 for the verb ACCÉDER, [au cache primaire] is an actant introduced by the preposition "à". But with other verbs, complements introduced by the same preposition correspond to circumstants (*e.g.*, [à ce stade] when linked to the verb ABANDONNER is a circumstant).

So, in order to identify actants and circumstants introduced by a preposition, we compute the probability P:[3]

$$P(\text{type} \mid \text{lexie}, \text{prep}, \text{nomprep}) = \frac{\#(\text{type}, \text{lexie}, \text{prep}, \text{nomprep})}{\#(\text{lexie}, \text{prep}, \text{nomprep})}$$

This probability calculates the number of times (frequency) that a given *lexie* (*i.e.* lexical unit) appears with a preposition as actant or circumstant. If its frequency as actant is higher, then we will decide it is an actant and not a circumstant.

2.1.3. Actant with the semantic role agent

In the first stages of the project, we annotate only actants and the feature assigns them a semantic role. The corresponding rules are shown in (9) and (10).

<*word, actant, OBJ*> → *patient* (9)

<*word, actant, SUJ*> → *{agent, instrument, cause, source…}* (10)

[3] The probability P is estimated by the ratio between the number of times the preposition as actant or circumstant appears with *lexie* and the number of times this *lexie* appears with or without this preposition.

Generally, we have seen in our corpus that the feature <Nom, OBJ> assigns a semantic role "patient".

In the case of (10), the feature <Nom, SUJ> cannot assign a semantic role. In many examples, this feature assigns agent for some units, instrument for other, etc. In order to identify this semantic role, we estimate the probability P:

$$P(\text{role} \mid \text{lexie}, FS, GS, \text{word}) = \frac{\#(\text{role}_i, \text{lexie}, FS, GS, \text{word})}{\#(\text{lexie}, FS, GS, \text{word})}$$

This probability calculates the number of times (frequency) that a given lexie (*i.e.* lexical unit) appears with a word (taking into account its part of speech and syntactic function) and a role as agent, patient, destination, etc. The most frequent role is selected.

2.2. Application of rules on an example

The sentence reproduced in Figure 3 has three participants indicated by the links at the bottom; the application of our rules to each participant is as shown in Figure 4.

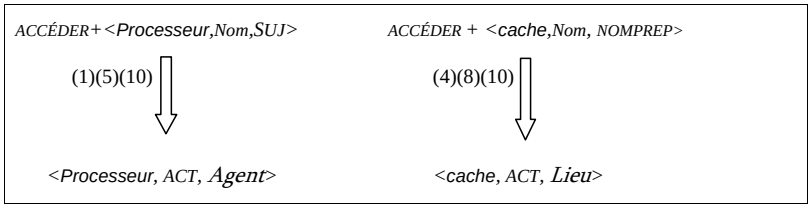


Figure 4. Application of rules to participants of sentence of Figure 3

In Figure 4, the feature of [processeur] is <*Nom,SUJ*>, the rule (1) and rule (5) are applied and the result is that [processeur] is an actant.[4] When this actant is identified, we must attribute to it a role. Then rule (10) is applied here and the role selected is Agent. The same work is done to the noun cache. The rules applied here are (4), (8) and (10) and selected the role location for this actant.

3. Conclusion

We have successfully developed a method for extracting rules to identify actants of French verbal lexical units. These rules extract features for assigning semantic roles like agent or patient. The application of these rules to more than two thousand manually annotated sentences produced results with an accuracy of 70% and a recall of 83%. In the future, other semantic roles such as destination, location or instrument will be considered.

[4] Rule (1) decides if processor is a participant of the verb "accéder" or not and rule (2) decides if it is an actant or a circumstant.

This automatic process should help alleviate the current annotation which is still carried out manually. Automatically annotated sentences will be submitted to terminologists who will then only have to edit the output. This work contributes to the design and availability of lexical and terminological resources containing rich linguistic information.

References

BAKER, C., FILLMORE, C. and LOWE, B. (1998). *The Berkeley FrameNet project*. Proceedings of Coling-ACL, vol. 1: 86-90.

BOURIGAULT, D., FABRE, C. and JACQUES, M.-P. (2005). L'analyseur syntaxique de corpus SYNTEX. *In Actes des 12e journées sur le Traitement Automatique des Langues Naturelles*. Dourdan.

FILLMORE, C.J. (ed.) (1982). Frame Semantics. Linguistics in the Morning Calm, In *Seoul International Conference on Linguistics (SICOL-'81')*: 111-137.

GILDEA, D. and HOCKENMAIER, J. (2003). Identifying semantic roles using combinatory categorial grammar. In *Proceedings of the EMNLP*, Sapporo.

GILDEA, D. and JURAFSKY, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, 28(3): 245-288.

GILDEA, D. and PALMER, M. (2002). The necessity of syntaxic parsing for predicate argument recognition. In *Proceedings of ACL'02*, Philadelphia.

MEL'ČUK, I. (2004). Actants in Semantic and Syntax. In *Actants in Semantics. Linguistics*, 42(2): 1-66.

RUPPENHOFER, J., ELLSWORTH, M., PETRUCK, R.L.M., JOHNSON C. and SCHEFFCZYK, J. (2006). *FrameNet II: Extended Theory and Practice* (http://framenet.icsi.berkeley.edu/book/book.html).

FrameNet. (http://framenet.icsi.berkeley.edu/).

SURDEANU, M., HARABAGIU, S., WILLIAMS, J. and AARSET, P. (2003). Using predicate-argument structures for information extraction. In *Proceedings of ACL'03:8-15*, Sapporo.

WordNet (http://wordnet.princeton.edu/).

Automated extraction of neologisms for lexicography

Jakob Halskov[1], Pia Jarvad[1]
Danish Language Council, Copenhagen

Abstract

The paper describes the implementation and evaluation of an automatic neologism detection prototype, the *Word trawler*. In two different experiments involving newspaper texts three detection techniques are tested, namely primitive filtering, statistical "weirdness" by comparison with a reference corpus predating the analysis corpus, and neology markers. It is found that a combination of these techniques results in the highest precision (approximately 40%). However, neology markers drastically reduce recall and should only be used when ample data is available. The authors finally suggest that diachronic frequency profiling could be used to further reduce system "noise", such as occasionalisms and spelling errors.

Keywords: neologisms, automatic, extraction, evaluation.

1. Introduction: What is a neologism?

Arriving at an operational definition for the concept of "neologism" is difficult. First of all, a neologism is new but relative to what? Secondly, it may be new but completely insignificant, for example a semantically transparent, trivial compound like "klimakonference" (*climate conference*). Thirdly, many candidate neologisms represent transient phenomena and are thus transient themselves, for example "klimakaravane" (*climate caravan*). This name was given to a bus touring Denmark in 2009 and distributing information on global warming. Such words will never reach the stage of "Lexikalisierung/Integration" (*cf.* Teubert 1997). Fourthly, candidate neologisms are really often occasionalisms or so-called nonce-formations (*cf.* Fischer 1998) or "lejlighedsdannelser" (Jarvad 1995). These are infrequent, often idiosyncratic and highly transient, expressions like "osteskuffe" (*cheese drawer*).

Given evidence that the new word or usage is becoming institutionalized, true neologisms, on the other hand, include:

(1) words referring to new objects or phenomena, for example "klimacertifikat" (*climate certificate*);

(2) new words replacing old ones, for example "sort" for "neger" (*black* replacing *negro*);

(3) semantic expansion of existing words, for example "blæksprutte" (*octopus*) in the sense "altmuligmand" (*odd job man*);

[1] {jhalskov,jarvad}@dsn.dk

(4) new valency of existing words, for example “dumpe en eksamen” replacing “dumpe **til** en eksamen” (*flunk an exam*);

(5) new multi-word expressions or phrases.

The key criterion for including a neologism in a dictionary of new words must necessarily be its assumed significance. As described above, significance is a composite measure consisting of multiple dimensions, including (assumed) transience, semantic transparency, cultural impact and so on.

2. Automatic extraction of neologisms: Word trawler architecture

The benefits of implementing a semi-automatic neologism extraction system are obvious. Manually monitoring the vast volumes of linguistic usage produced in modern information society is impossible, and thus manual extraction work is bound to be biased. While assessing the significance of a candidate neologism may often be a fairly easy task for a trained human (linguist), it is by no means a trivial problem for a computer, however.

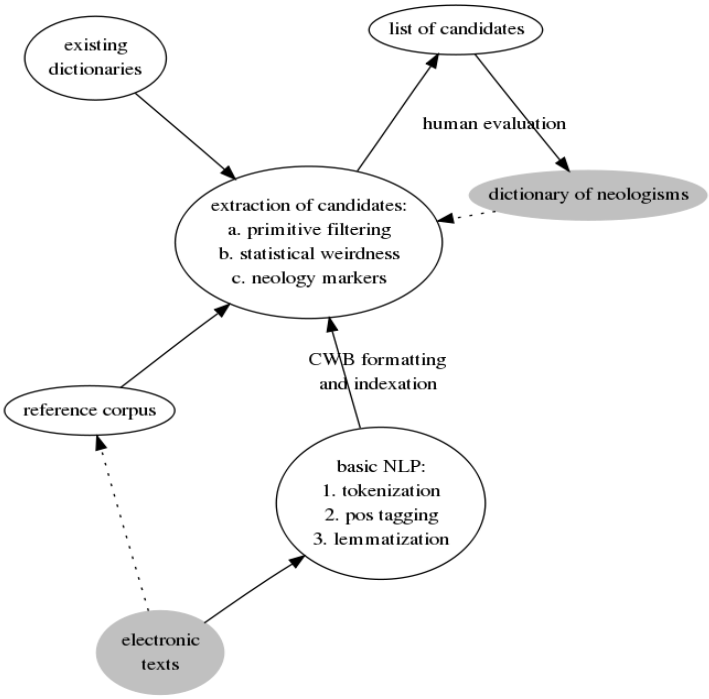


Figure 1. The architecture of the Word trawler extraction system

As shown in Figure 1, we have implemented a semi-automatic extraction prototype, the “Ordtrawler” (*Word trawler*), whose techniques are partly inspired by the findings

of the Analysis and Prediction of Innovation in the Lexicon (APRIL) project headed by Antoinette Renouf.[2]

As indicated by the grey circles the starting point is a great number of electronic texts, and the end product is a dictionary of neologisms. The texts are subjected to basic NLP and candidate neologisms are subsequently extracted by means of a combination of three different techniques, namely primitive filtering (which relies on existing dictionaries), statistical weirdness (which relies on word frequencies from a reference corpus predating the analysis corpus) and neology markers like "såkaldt" (*so-called*).

3. Extraction techniques

Primitive filtering is executed using the existing dictionaries (1-3) and corpora (4-5) listed in Table 1.

	Filter	**#tokens**	**#lemmas**	**#types**
1	"Retskrivningsordbogen" 2001	n/a	64,038	399,062
2	In "Den Danske Ordbog", but not in 1	n/a	34,960	n/a
3	In "Ordsamlingen" but not in 1-2	n/a	221,679	n/a
4	In "Korpus 90", but not in 1-3	28 M	?	124,585
5	In "Korpus 2000", but not in 1-4	28 M	?	436,004
Total			**?**	**1,216,290**

Table 1. Primitive filtering of candidates using existing dictionaries and corpora

"Ordsamlingen" is not really a dictionary but a large, and partly digitized, collection of previously recorded neologisms (in context). The largest Danish corpus of general language, "Korpus DK",[3] is used as the reference corpus. It is the union of "Korpus 90" (covering the years 1988-1992) and "Korpus 2000" (covering the years 1998-2002). Statistical "weirdness" (Ahmad 1993) is measured with log-odds ratio as described in *e.g.* Evert (2004). This association measure favours rare events, and neologisms are typically rare events in a corpus as was found in the APRIL project. As for the neology markers, only two different markers are used in the experiments, namely "såkaldt" and "såkaldte" (so-called).

4. Experiment 1: evaluation of recall

In this experiment all neologisms present in a small analysis corpus comprising 177 short newspaper articles (c. 75,000 tokens and 14,000 types) were manually tagged. The resulting gold standard contains 252 neologisms including 33 multiword

[2] See http://gow.epsrc.ac.uk/ViewGrant.aspx?GrantRef=GR/L08243/01
[3] http://ordnet.dk/korpusdk_en/front-page/view?set_language=en

expressions and 11 semantic extensions of existing lemmas. Multiword expressions and semantic extensions are excluded from the evaluation since they are not presently handled by the system.

4.1. Primitive filtering and neology markers

Table 2 summarizes the performance of the Word trawler using primitive filtering or markers. The best balance between precision and recall is obtained using all filters and excluding all proper nouns from the result. This configuration yields 589 candidate neologisms of which 124 (21%) are found in the gold standard. The numbers also reveal that eliminating all candidates which occur in "Korpus 2000" boosts precision but reduces recall. The reduction in recall can be explained by the fact that "Korpus 2000", although it represents texts from the years 1998-2002, may still contain words which can be considered new in 2010. Finally, neology markers give a high precision, but an extremely poor recall.

	#candidates	**F-score**	**Recall**	**Precision**
Human	208	1	100%	100%
Word trawler (incl. proper nouns)	1061	0.22	69%	13%
Word trawler (excl. proper nouns)	589	**0.31**	60%	21%
Word trawler (incl. proper nouns and deactivating the "Korpus 2000" filter)	1498	0.20	**84%**	12%
Word trawler (excl. proper nouns and deactivating the "Korpus 2000" filter)	878	0.28	73%	17%
Word trawler (neology markers)	15	0.05	2.9%	**40%**

Table 2. Word trawler performance (primitive filtering)

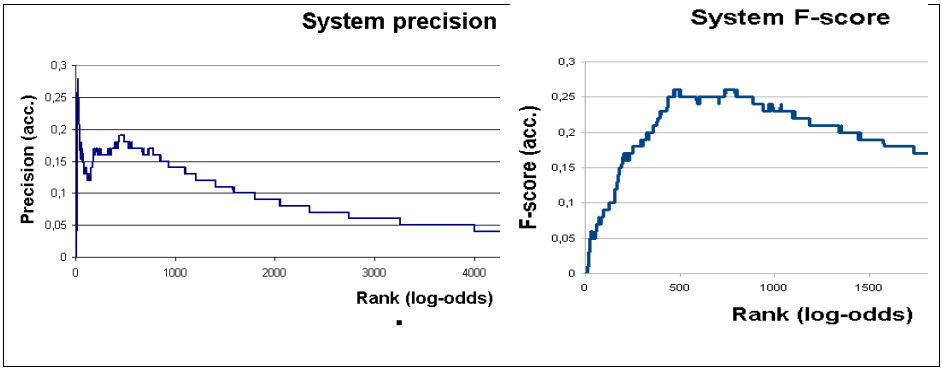


Figure 2. Word trawler performance (statistical ranking)

4.2. Statistical ranking by "weirdness"

Using the slightly more sophisticated technique of ranking all word forms in the analysis corpus by their statistical weirdness, yields precision and F-scores as depicted in the plots of Figure 2. With this technique semantic extensions are also detected.

5. Experiment 2: evaluation of precision

In this experiment the analysis corpus is several magnitudes larger than the previous one. It comprises one year of newspaper articles from the same source but totalling 96.7 million tokens. Since neology markers proved to be the most precise extraction technique, they were combined with the primitive filters described above in an attempt to boost system precision as much as possible while ignoring recall.

The extraction produced 1,784 candidates from which the most frequent 200 and the least frequent 200 candidates were sampled and manually evaluated by two human judges. With an inter-annotator agreement of 84.4% the two judges unanimously announced 152 of the 400 candidates, almost 40%, to be true and significant neologisms.

5.1. Classification of noise

The 248 candidates which were not accepted as neologisms by both human judges were grouped into nine different noise categories. These categories are presented in Table 3. Each candidate may instantiate more than one type of noise.

Noise category	Example	Number	Ratio
Inflected forms	"undersøgelseskommissioner" (inquiry committees)	83	28.8%
Transparent compounds and occasionalisms	"Forskningskvalitet" (research quality), "fodboldekspert" (soccer expert), "pizzabande" (pizza gang)	81	28.1%
LSP	"kapillærvirkning" (capillary action)	44	15.3%
Spelling mistakes		25	8.7%
Filter errors	"nummerportering" (number porting)	14	4.9%
Code shifts	"Surge", "caucus", "caviats", "stâuerna"	13	4.5%
Proper nouns	"JPMorgan", "SEA-Games", "TMM"	12	4.2%
Old expressions	"epidemihus" (historical institution, a precursor of the modern hospital)	11	3.8%
NP fragments	"parkér (og rejs)" (park and go)	5	1.7%
Total		288	100%

Table 3. Noise classification

The main type of noise is inflected forms and is caused by the system's primitive lemmatizer which has trouble handling out-of-vocabulary items. An equally predominant problem is transparent compounds and occasionalisms. These are hard even for humans to distinguish from significant neologisms, so this is not surprising. Language for Special Purposes (LSP) and spelling mistakes are also important sources of noise. Making a machine distinguish LSP from general language neologisms is also very challenging. The only obvious way of reducing such types of noise is arguably by generating diachronic frequency distribution profiles for all candidates and eliminating candidates whose profiles do not resemble the prototypical pattern. Finally, proper nouns and NP fragments can be avoided by implementing more sophisticated NLP such as a Named Entity Recognizer and a phrase chunker.

6. Conclusions and future work

The evaluation of the Word trawler system illustrates that automatic detection of significant neologisms in natural language text is non-trivial. Although the performance of the Word trawler system could presumably be improved by equipping it with more sophisticated NLP, its precision is unlikely to exceed 40% to 50%. And in order to achieve even this level of precision, recall must be sacrificed and large amounts of text must be processed. However, the key strength of automated neologism detection is that neither processing time nor subjectivities or idiosyncracies affect the extraction process.

As for future work the authors will focus their efforts on introducing diachronic frequency profiling to reduce noise like occasionalisms, spelling errors and technical terms.

References

AHMAD, K. (1993). Pragmatics of specialist terms: The acquisition and representation of terminology. In P. Steffens (ed.). *Machine Translation and the Lexicon. Third International EAMT Workshop*. *Lecture Notes in Computer Science*, vol. 898. Heidelberg: Springer: 51-76.

EVERT, S. (2004). *The Statistics of Word Co-occurrences: Word Pairs and Collocations*. PhD dissertation. University of Stuttgart.

FISCHER, R. (1998). *Lexical Change in Present-Day English – A Corpus-Based Study of the Motivation, Institutionalization, and Productivity of Creative Neologisms.* Tübingen: Günter Narr Verlag.

JARVAD, P. (1995). *Nye ord i dansk – hvorfor og hvordan?* Copenhagen: Gyldendal.

TEUBERT, W. (1997). *Neologie und Korpus*. Tübingen: Günter Narr Verlag.

Tickbox Lexicography

Adam Kilgarriff[1], Vojtěch Kovář[2], Pavel Rychlý[2]
Lexical Computing Ltd, Masaryk University

Abstract

Corpus lexicography involves, first, an analysis of a word, and then, copying of collocations and examples from corpus to dictionary. In a large project, there are hundreds of thousands of items to be copied across. Most modern lexicography takes place on a computer, with both corpus and dictionary editor being software applications and the copying done by re-keying or copy-and-paste. This can be inconvenient and time-consuming. We present a more efficient approach, where the user selects collocations and examples by clicking on tickboxes and the material is automatically structured and formatted according to the particular dictionary's requirements, ready for pasting into the dictionary editor.

Keywords: corpus lexicography, dictionary editing.

1. Corpus Lexicography

In corpus lexicography we:

- identify the senses for the word

and then, for each sense:

- identify the key patterns, collocations and phrases
- find example sentences.

This process is the core of the lexicography for a language. Once it has been completed for the full vocabulary, the resulting database is a base analysis of the language which will serve for the development of a range of dictionaries, monolingual and bilingual (where the language analysed is the source language, and the analysis will form the basis whatever the target language) (Atkins 1994; Atkins and Rundell 2008: 97-101).

In a large project, there are hundreds of thousands of items to be copied across from corpus to dictionary editor. Most modern lexicography takes place on a computer, with both corpus and dictionary editor being software applications and the copying done by re-keying or copy-and-paste. This can be inconvenient and time-consuming. We present a more efficient approach, where the user selects collocations and examples by clicking on tickboxes and the material is automatically structured and formatted

[1] Lexical Computing Ltd, UK, adam@lexmasterclass.com
[2] Masaryk University, Brno, Czech Republic, {xkovar3,pary}@fi.muni.cz

according to the particular dictionary's requirements, ready for pasting into the dictionary editor.

Following a brief note on our corpus application, the Sketch Engine (Kilgarriff *et al.* 2004), we describe TickBox Lexicography (TBL). We then give an account of how it is being used in two large-scale projects, at Macmillan Publishing in the UK and at the Institute for Dutch Lexicology (INL) in the Netherlands.

2. The Sketch Engine

The Sketch Engine is a leading corpus query tool, in daily use for lexicography at publishing houses such as Oxford University Press, Cambridge University Press, Collins and Macmillan in the UK, INL in the Netherlands and Cornelsen in Germany, and for language research and teaching at a number of universities worldwide. It operates as a ready-to-use online service, with large corpora available to all customers for most of the world's major languages.

There are two functions which the Sketch Engine offers which support the process:

- 'word sketches', one-page summaries of the key collocations (and sometimes phrases) for the word, in a table organised by grammatical relations (*e.g. object, modifier, modified*) (*cf.* Figure 1).
- the 'Good Dictionary Example eXtractor', GDEX, a function for finding good dictionary examples (Kilgarriff *et al*. 2008).

GDEX is far from perfect and we cannot assume that the 'best example' according to GDEX is good enough to go straight into a printed dictionary. But it does greatly improve the chances that the lexicographer will find a useable sentence (often needing some further editing) amongst the first few concordance lines for a collocation.

3. Tickbox Lexicography

Tickbox Lexicography is a variety of corpus lexicography with intensive computational support in which the lexicographer selects aspects of the automated analysis of the word for inclusion in the dictionary by ticking boxes, and then pastes their selections into the editing interface.

The process is as follows:

- the lexicographer sees a version of the word sketch with tickboxes beside each collocation.
- for each sense and each grammatical relation, they tick the collocations they want in the dictionary (see Figure 2).
- they click a 'next' button.

- they then see, for each collocation they have ticked, a choice of six (by default) corpus example sentences, chosen by GDEX, each with a tickbox beside it: they tick the ones they like (see Figure 3).
- they tick a “copy to clipboard” button.

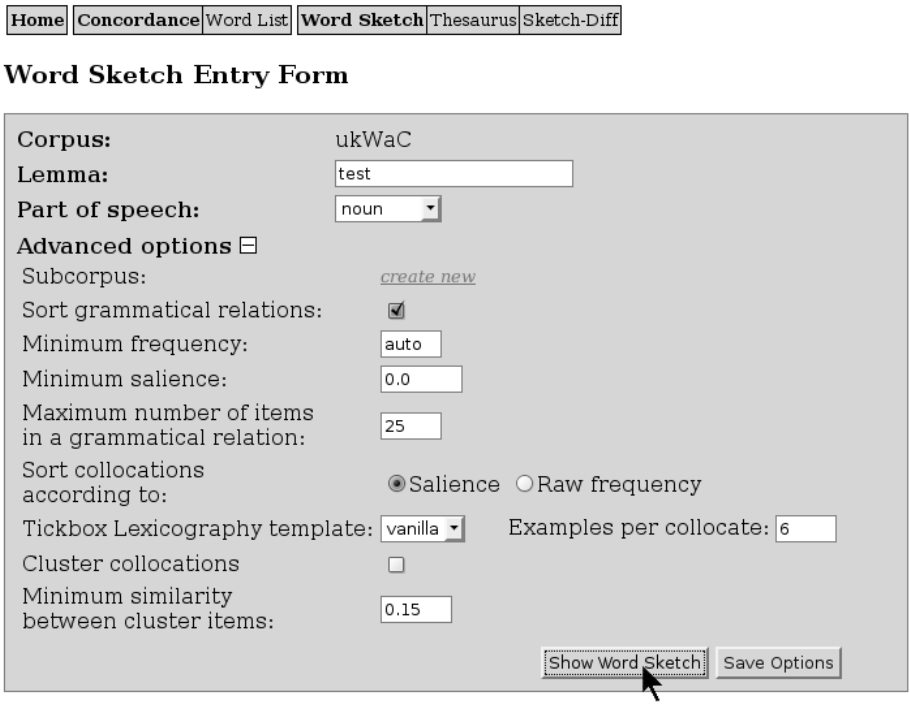


Figure 1. Word Sketch Entry Form

Each target dictionary has its own TBL application, based on the DTD (for XML-based systems) or field names used in the dictionary. The system then copies the collocations and examples, embedded in an XML structure as required by the user’s dictionary-editing system and target dictionary, onto the clipboard. (The XML fragment for the example above, with a ‘vanilla’ DTD, is shown in Figure 4.). The lexicographer can then paste the structure into the dictionary editing system. For this we use the operating system’s clipboard functions. While this is not, in computer science terms, the most elegant technique, we have found it to be the most convenient, and widely-applicable method for transferring data between programs on the same computer. It has the great advantage that all users know and understand it.

Thus, TBL models and streamlines the process of getting corpus data out of the corpus system and into the dictionary editing system.

Figure 2. Word sketch for the English noun ‘test’*, data from the UKWaC corpus*

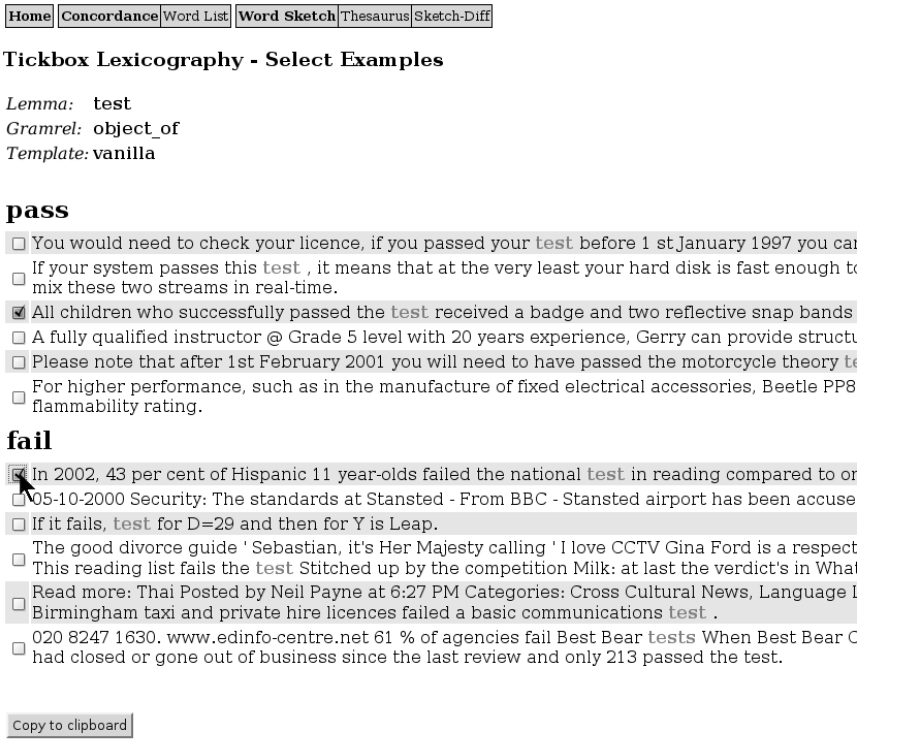


Figure 3. GDEX examples for each selected collocate.

```
- <entry>
    <keyword>test</keyword>
  - <gramrel>
      <grname>object_of</grname>
    - <collocation>
        <collo>pass</collo>
      - <example>
          All children who successfully passed the
          <b>test</b>
           received a badge and two reflective snap bands on the day and a c
        </example>
      </collocation>
    - <collocation>
        <collo>fail</collo>
      - <example>
          In 2002, 43 per cent of Hispanic 11 year-olds failed the national
          <b>test</b>
           in reading compared to only 16 per cent of white children.
        </example>
      </collocation>
    </gramrel>
  </entry>
```

Figure 4. An XML entry draft, as copied to the clipboard

4. Projects

At time of writing there are two large-scale dictionary projects where TBL is in daily use: the *Macmillan Collocations Dictionary* and the *Algemeen Nederlands Woordenboek*.

4.1. Macmillan Collocations Dictionary

At Macmillan Publishing, the *Macmillan Collocations Dictionary* (MCD) is in preparation. MCD starts from MEDAL (2007), and provides a full account of the collocations of the core senses of around 4,000 common and highly 'collocational' words (Kilgarriff 2006). As in word sketches (and in other collocations dictionaries such as Oxford's (OCD 2002, 2009), collocations are organised according to the grammatical relations. Some collocations are illustrated with examples in the paper book; all have examples available by mouse-click in online and other electronic versions.

To set up TBL for MCD, we first developed customised word sketches in which the grammatical relations were those to be used in MCD. This required work on the underlying part-of-speech tagging and grammatical-relation-finding software. (The parsing, to identify grammatical relations, uses regular expressions over part-of-speech tags and is built in to the Sketch Engine: see Kilgarriff *et al.* 2004.) GDEX was also customised, with the incorporation of a long list of 'stop' words, to minimise the chances that GDEX would select examples containing offensive material.

In the first trials, lexicographers selected all the example sentences (typically six per collocate) that were to be used in the electronic version of MCD, but this proved too slow. We changed to a strategy where only the examples which are to appear in the book are selected by lexicographers. For all others, GDEX will be trusted to deliver good examples. (The manually-selected items will be edited as necessary by lexicographers, whereas the others will be full and unedited corpus sentences.) These sentences will be selected in a batch process after the main phase of the lexicography is complete, as this will reduce the volume of data to be handled by the clipboard and the dictionary editing system, and will allow us to use a new version of GDEX. We anticipate using the experience gathered during the project to fine-tune GDEX according to Macmillan's preferences and observations, and this can only take place at or near the end of the project.

MEDAL 2007 already contains 1000 'collocation boxes' for word senses of common words, with collocations classified according to grammatical relations, and further collocations in bold in regular entries. It was desirable to carry them across into MCD, in a way which integrated with MCD lexicography. To this end we:

- analysed MEDAL to find all collocations, either in collocation boxes or shown in bold within regular entries
- identified the grammatical relation they stood in to the headword
- checked to see if they were already in the word sketch:

- if they were (as they usually were), colour them red (in the word sketch) and pre-tick the tickbox, as they will almost always be wanted in MCD
- if they were not, add them in (in red), with links to their corpus instances and pre-ticked tickboxes.

The dictionary editing software used for MCD accepts XML pasted from the clipboard. This means that, once the lexicographer has:

- called up the customised word sketch for the headword,
- selected the grammatical relation,
- selected collocates,
- selected examples for the paper dictionary,

… they click a 'copy to clipboard' button, and then paste the material (using standard CTRL-V) into the dictionary entry.

4.2. Algemeen Nederlands Woordenboek (ANW)

At the Institute for Dutch Lexicology (INL), the *Algemeen Nederlands Woordenboek* (General Dutch Dictionary, ANW) is a large Dutch dictionary project running from 2001 til 2018. The project has been using the Sketch Engine for corpus access since 2007 (for background and a full account, see Tiberius and Kilgarriff 2009).

Within the ANW dictionary project, example sentences are gathered together with bibliographic information for each citation.

TBL offers an architecture for efficiently collecting information from the corpus and packaging it for insertion into the dictionary database. While the system had been designed with linguistic information in mind, it was readily adjusted for the ANW dictionary project to gather and insert bibliographic information as well. We developed a TBL installation with a specific template for INL where, when the lexicographer ticks the example, not only the example but also the title, author, publisher and date of its source are assembled in an XML fragment and placed on the clipboard. (We also made it possible to select multiple examples at a time, as this fitted the way that INL lexicographers worked.) The dictionary editing software (Niestadt 2009) was customised to interpret these XML structures so, when the user pastes the example from the clipboard into the editor, the different components of the reference are placed in the appropriate database fields with a single mouse-click.

5. Conclusion

We have shown how TBL works in the general case, and how it has been used in two large projects. We believe TBL has great potential for both streamlining corpus lexicography and making it more accountable to the corpus.

References

ATKINS, B.T.S. (1994). A corpus-based dictionary. In *Oxford-Hachette English-French Dictionary* (Introductory section). Oxford: Oxford University Press: xix-xxxii.

ATKINS, B.T.S. and RUNDELL, M. (2008) *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.

KILGARRIFF, A. (2006). Collocationality (and how to measure it). In *Euralex Proceedings.* Torino.

KILGARRIFF, A., RYCHLÝ, P., SMRŽ, P. and TUGWELL, D. (2004) The Sketch Engine. In *Euralex Proceedings.* Lorient: 105-116.

KILGARRIFF, A., HUSÁK, M., MCADAM, K., RUNDELL, M. and RYCHLÝ, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Euralex Proceedings*, Barcelona.

MEDAL (2007). *Macmillan English Dictionary for Advanced Learners*. Second edition. Ed. by M. Rundell. Oxford.

NIESTADT, J. (2009). De ANW-artikeleditor: software als strategie. In E. Beijk *et al.* (eds). *Fons Verborum: Feestbundel Fons Moerdijk.* Amsterdam: Gopher: 215-222.

OCD (2009 [2002]). *Oxford Collocations Dictionary*. Oxford.

TIBERIUS, C. and KILGARRIFF, A. (2009). The Sketch Engine for Dutch with the ANW corpus. In E. Beijk *et al.* (eds). *Fons Verborum: Feestbundel Fons Moerdijk*. Amsterdam: Gopher BV.: 237-255.

ABBYY Lingvo electronic dictionary platform and Lingvo Content dictionary writing system

Vera Kuzmina[1], Anna Rylova[2]
ABBYY

Abstract

The art and craft of lexicography – one of the oldest sciences in the world – has been developing and improving for many years. But the computer technologies started to serve lexicography only about 50 years ago. Data creation, manipulation, storage and usage started being carried out on computers. In this situation dictionaries were strongly involved in the process – they started being created and modified on computers with the result of the lexicographer's work being read by the users on computers as well as on handheld devices and mobile phones. In this paper we give an overview of electronic lexicography in Russia and describe the technologies of the ABBYY company, which developed the Lingvo electronic dictionary 20 years ago and then started creating a dictionary writing system, an online portal and dictionaries for mobile phones. Through all these years the company has not only been developing software products, but also conducting research in linguistics, semantics, syntax and lexicography, practicing at the same time dictionary creation – more than 40 dictionaries were created by ABBYY's lexicographers in collaboration with external compilers. We describe the main features of the ABBYY Lingvo electronic dictionary and the ABBYY Lingvo Content dictionary writing system and show prospects for their future development in mutual interchange between dictionary developing, software making, linguistic research and study of the dictionary users' needs.

Keywords: electronic dictionary, online dictionary, mobile dictionary, dictionary writing system.

1. ABBYY Lingvo electronic dictionary

More then 20 years ago, the Lingvo electronic dictionary was created in Russia. Making electronic dictionaries and releasing them as software products was quite a new phenomenon in the era when electronic lexicography itself was just starting its development. One of the important ideas since Lingvo dictionary moved from DOS to Windows platform (this happened in 1993) was that electronic dictionary users can create their own dictionaries in Lingvo format using a special markup language. For that purpose, software developers included a special compiling program in the electronic dictionary. The program could transform user's dictionaries into electronic dictionaries in Lingvo format.

[1] Head of Strategic Partnership Group, Vera_K@abbyy.com
[2] Product Analyst, Russian State University for the Humanities, Anna_Ry@abbyy.com

1.1. A dictionary compiling tool for the Lingvo community

The dictionary compiling program and the markup language were not yet a dictionary writing system, but a reliable, useful and fully fledged instrument for creating electronic dictionaries. It is worth mentioning that this feature (a possibility of creating your own dictionary, viewing it through the Lingvo electronic dictionary software and sharing it with other users) made the Lingvo dictionary very popular and formed a big community of dictionary makers who became Lingvo fans. This community of people who create dictionaries and are generally interested in lexicography, linguistics and translation, is growing bigger and connects different people interested in these areas.

Lingvo dictionary software, which had such features as integration with Windows programs (a possibility to translate words from Windows applications immediately using hot keys), the possibility of switching between different dictionaries included in one Lingvo pack, seeing the translation of a word in universal and special dictionaries and using cross-references, became an essential tool for dictionary makers and users. All these features were included in the Lingvo dictionary by 1993.

A professional DWS, which should consist of a text-editing interface, a database and a set of administrative tools (Atkins and Rundell 2008) was later developed by ABBYY.

1.2. Sophisticated search facility

As Grefenstette (1998) states, "It is clear the one of the things that a computer can do well is treat a large amount of data" and this is true for electronic dictionaries as well. In 1997, the 5th version of the Lingvo electronic dictionary was released and it included 400,000 dictionary entries. New means of working with such a large amount of data were needed. That year, a new edition of Lingvo included a morphological engine and had new search capabilities. Since then dictionary users have been able to translate a word with Lingvo even if it is not in the initial form. The full-text search function that was added to the application enables the users to find real usage examples, collocations, or typical word combinations shown in context.

In the modern version of Lingvo searches are carried out in the following entry sections: headwords, translation equivalents, examples, and comments. At the same time, the "wildcard" search option and the ability to search in multiple languages at a time enables lexicographers to find paronyms and related words, or trace the evolution of the meanings of borrowed words, for example the word "Handy" in German, which originally meant "convenient to handle or use; useful[3]" in English.

2. Accessing dictionary content via Lingvo Platform

Since that time the electronic dictionary has been developed, new functions were added to the software, but at the same time, the application of electronic dictionaries broadened. Lingvo became not only a desktop application, but also a mobile and

[3] Oxford Dictionary of English, Revised Edition. © Oxford University Press 2005.

online one. This led to the integration of several new components into the platform. These products are connected to the company's data centre and provide quick and easy access to the dictionary and reference content which is stored on it. This client-server model is the ideal tool to offer more content for the end user upon his real requests for it, and thus ABBYY Lingvo has become a platform for collaboration with different publishers and authors who are interested in providing their content to end users of electronic dictionaries by different means: PC, Internet or mobile.

Throughout the 20 years of lexicographic software development, a lot of attention has been paid to the product's user friendliness and graphical interface as well as coverage of as many usage scenarios as possible: checking translations, finding appropriate translation equivalents, looking up the meaning of a word in a particular subject domain, checking spelling, learning new words, and even enjoying oneself while browsing the dictionaries.

3. Writing dictionaries with ABBYY Lingvo Content

The electronic dictionary software development was combined in ABBYY with constant research in linguistics, syntax, semantics, lexicography and active practical work on dictionary making – more than 40 dictionaries were created by ABBYY's lexicographers in collaboration with external compilers. The markup language and compilation tool developed in the beginning of the 1990s were not enough to support big lexicographic projects. So the company started developing a dictionary writing system to satisfy the needs of in-house lexicographers and all the individual authors who created dictionaries in Lingvo format.

ABBYY created its own dictionary writing system, the ABBYY Lingvo Content, which became for its users a tool for creating dictionaries from scratch, updating or supplementing existing dictionaries, exporting data to various formats for subsequent publication on paper or in an electronic format (including the formats used by the PC and mobile versions of ABBYY Lingvo and online and intranet dictionaries).

The dictionary writing system is based on a client-server architecture and supports multi-user online and offline work. Data is stored in an XML-based format on a database server. Unicode characters are supported (*e.g.* phonetic transcription symbols and Pinyin are displayed correctly).

3.1. Short overview of ABBYY Lingvo Content's main features

The ABBYY Lingvo Content dictionary writing system is a system developed by lexicographers for lexicographers. With the years of usage, the system changed, and new functions were added, but the main principle remained the same – the lexicographer does not necessarily need any special computer skills to work with the system and can be as unassisted in using it as is possible.

The main features of the system include:

- easy working with entry structure;
- automatic renumbering of entry elements;
- automatic cross-references update;
- spell-checking the text of the entries and validating their structure;
- entry version history and comparison with visual mark-up of changes;
- user-friendly entry filtration system (no special query language – only checking boxes);
- possibility of working with many dictionaries (2 and more) in one window, editing their entries simultaneously;
- tool for dictionary comparison and merging embedded in the interface;
- workflow and dictionary project management tools.

3.2. Using the filtration tool

Lexicographers can get any subset of entries from the dictionary using a sophisticated filtration tool that does not require any knowledge of special query language. Entries with any specific parameters or words in any entry field can be found easily by checking boxes in the filtration dialogue. The filtration tool suggests multiple parameters for filtration – find entries with specified label, specified word in any field, any number of senses and many other different possibilities.

The filtering results can be assigned to any user, saved as a separate dictionary or a batch within the main dictionary. They can also be compared and merged with other dictionaries or batches.

3.3. Merging and comparing dictionaries, working with dictionary groups

Several dictionaries can easily be compared using an intuitive interface tool. Comparison results can be saved as a separate batch or added to any dictionary. For each entry its source (dictionary name) will be indicated. It is also possible to merge different dictionaries and batches that can be viewed and edited simultaneously in one window with the source dictionary name indicated.

3.4. Managing your dictionary project

All the entry versions are saved, and the lexicographer can compare any two entry versions at any time and see what was added, deleted or changed in the entry. If the lexicographer does not like the changes, it is possible to reject them and go back to a previous version of the entry.

In a multi-user lexicographic project, it is possible to get information at any time about who did what (added entry, deleted entry, updated entry, changed headword etc.) and

when. Statistics for any period of time are available. Full statistics about the number of entries, translations, examples and other elements in the dictionary are also available.

The system supports multi-user access from remote computers with automatic locking of entries that are being worked on. A workflow management system is used for planning, task distribution, and control.

3.5. Publishing dictionaries

Both the ABBYY Lingvo Content DWS and platform are useful lexicographic tools enabling authors to create and update dictionaries and easily export them from the DWS into the end-user application, which is available in online, mobile and desktop versions. If a dictionary is to be printed on paper, it is exported into a publishing system via XML, RTF or DOCX file generated from the system.

4. Some future prospects

In the new information era, dictionary projects are growing bigger, involving a great amount of language material and providing dictionary users with different kinds of information. Therefore, the future of big dictionary projects is in collaborative work where professional lexicographers and dictionary users work together in close collaboration, both having the tools to work on dictionary projects. The dictionary creation tools must be client-server and available online for different people inside the dictionary making community.

We strongly believe that with the development of dictionary writing and dictionary usage new tasks and challenges will appear. This will move the dictionary writing and dictionary viewing tools further.

References

ATKINS, B.T.S. and RUNDELL, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

GREFENSTETTE, G. (1998). The future of linguistics and lexicographers: will there be lexicographers in the Year 3000? In Th. Fontenelle *et al.* (eds.) *Proceedings of the Eighth EURALEX Congress*. Liege: University of Liege: 25-41.

http://www.abbyy.com/company/

http://www.lingvo.com

Dictionary management system for bilingual dictionaries

Margit Langemets[1], Andres Loopmann[1], Ülle Viks[1]
Institute of the Estonian Language (Tallinn, Estonia)

Abstract

Our project focuses on a web-based dictionary management system for bilingual dictionaries. The system is designed for different users (lexicographers, language learners, translators, etc.) who desire to compile their own bilingual (primarily Estonian–other) dictionary via the web. The system enables users to design their own dictionary as they need it, choosing the languages, modifying the entry layout and structure, and using a ready-made description of the source language. The main components of the management system are: (1) EELex system of dictionary administration, (2) an Estonian–X dictionary database, and (3) the management system interface.

Keywords: web-based dictionary management, bilingual dictionaries, automatic morphology.

1. Introduction

The open world of today experiences a growing need for translation dictionaries. Our dictionary management system for bilingual dictionaries is meant to facilitate lexicographer's work and, at the same time, enhance its quality. We believe that dictionary-makers should be relieved from technical problems, so that they could focus on the most important task of entry compilation. Our aim is a web-based universal dictionary writing system for compilation, editing and publication of a bilingual dictionary, with an interface enabling users to format the dictionary as they like it, choosing the source and target languages, and modifying the entry layout and structure. An additional asset is an Estonian-X dictionary database and morphological synthesis of Estonian wordforms.

2. ΞΞLex system of dictionary administration

The **ΞΞLex** (= EELex) system of dictionary administration is a web-based lexicographer's workbench integrating various language technological tools: linguistic software and language resources (see Langemets *et al.* 2006). The main features of EELex are: Unicode support, XML databases and schemas, XSL transformations for generating different views (XML view, Edit view, Layout view), click-to-edit, structural queries and sorting of query results, export to the MS Word layout format, team work option (with different levels of user rights), various tools for entry and

[1] {margit@eki.ee,andres.loopmann,ylle}@eki.ee

dictionary editing, *e.g.* menu compiler, XML file generator, etc. (*cf.* Joffe *et al.* 2008, Mangeot 2006).

The EELex system of dictionary administration has been created at the Institute of the Estonian Language under the project called “Lexicographer’s workbench”. At present, the system involves about twenty dictionaries, of which five have been completed (published or finished), eleven are being edited, and two existing dictionaries (in another format) are being prepared for transferring to an EELex format.

Figure 1 shows the EELex editing window (language option of the user’s interface: English).

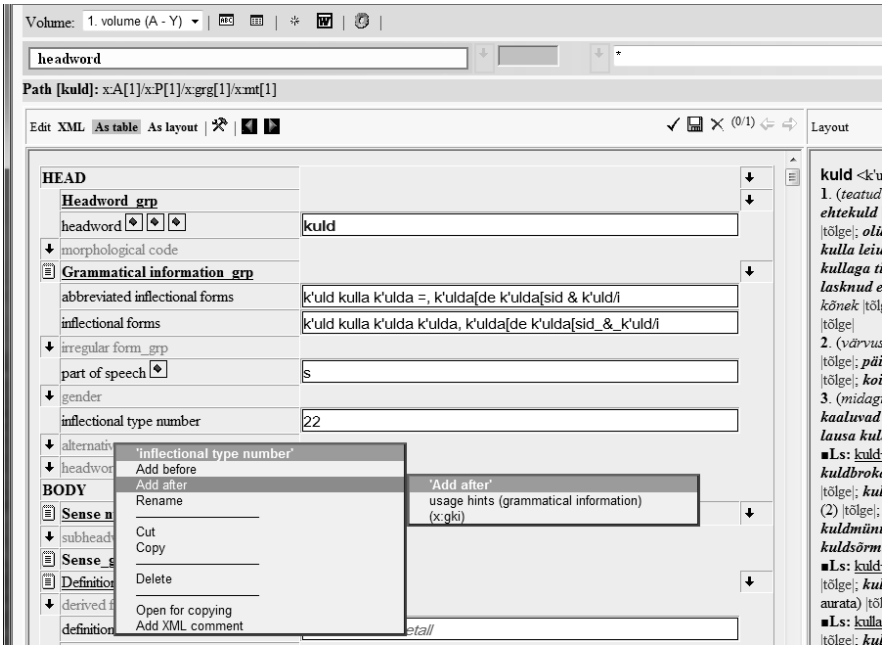


Figure 1. Editing window

3. Estonian–X dictionary database

The main EELex application is the Estonian–X dictionary database (= EXDD), which is compiled by means of EELex to provide a core for the new bilingual dictionaries to be produced by the system. The EXDD contains only source language (Estonian) data: entry word, grammatical information, explanations, labels, usage examples, compound words, etc. The data of the target language (X) – translation equivalents with the rest of the necessary information – will be supplied by the user compiling the new dictionary.

The EXDD source material comes from a voluminous (c. 80,000 entries) Estonian-Russian dictionary (Liiv *et al.* 1997-2009), plus some material from other dictionaries. The presentation of the material has been tailored for the EXDD, while an essential part of the entries (c. 40,000) have been subjected to a detailed editing process, using a standard presentation for sense division and homonyms, as well as for compound words (headword or example), cross-referencing, labelling (usage information, domains), etc. According to the type of dictionary required, three standards of morphological description have been developed to facilitate the presentation of Estonian morphology for non-native users who might otherwise be taken aback by the great number of inflected forms and extensive variation of morphological units.

Figure 2 presents an example (the entry *kuld* ‘gold’) from the EXDD in layout format. Note that the system presents a preliminary “standard” form of the entry, of which the lexicographer can modify, if necessary, both the content and the form.

(1) **kuld** <k'uld kulla k'ulda k'ulda, k'ulda[de k'ulda[sid_&_k'uld/i **22** ***S***>
1 (*teatud väärismetall* 'a yellow precious metal'***)*** |TE| ♦ ***puhas kuld*** |TE|; ***ehtekuld*** |TE|; ***kullast ehted*** |TE|; ***kulda pesema*** |TE|; ***lõpetasin keskkooli kullaga*** *kõnek* |TE|
2 (*värvuselt ja läikelt kulla sarnane* 'colour resembling gold') |TE| ♦ ***päikesekuld*** |TE|; ***sügiskuld*** |TE|; ***kased puistavad juba kulda*** |TE|
3 (*midagi väärtuslikku ja head* 'sth highly respected') |TE| ♦ ***tema nõuanded on kulda väärt*** |TE|; ***nendel sõnadel on kulla kaal v hind*** |TE|

(2) ■**Ls:** kuld+ (*kullast, kullatud* 'made of gold') ♦ ***kuldbrokaat*** *tekst* |TE|; ***kuldmedal*** |TE|; ***kuldmünt*** |TE|; ***kuldsõrmus*** |TE|; ***kuldvillak*** *müt* |TE|
■**Ls:** kuld+ (*kulla värvi* 'lustrous yellow') ♦ ***kuldblond*** |TE|; ***kuldjuukseline*** |TE|; ***kuldkollane*** |TE|; ***kuldpõrnikas*** *zool* (Cetonia aurata) |TE|
■**Ls:** kulla+ ♦ ***kullaauk*** *kõnek, piltl* |TE|; ***kullafond*** (1) *maj* |TE|, (2) *piltl* |TE|; ***kullakang*** |TE|; ***kullaketraja*** |TE|; ***kullaliiv*** |TE|; ***kullaläige*** |TE|; ***kullamäed*** *piltl* |TE|; ***kullaotsija*** |TE|; ***kullapalavik*** *piltl* |TE|; ***kullaproov*** |TE|; ***kullasoon*** |TE|; ***kullastandard*** *maj* |TE|; ***kullatera*** |TE|

Figure 2. The dictionary entry kuld (‘*gold’*)

Section (1) includes the main part of the entry: the headword with a full morphological description (all basic inflectional forms, the inflectional type number, part of speech). The numbers differentiate between senses (and explanations), while each sense is followed by usage examples. Section (2) contains a separate block of compounds associated with the headword, classified according to the grammatical form and meaning of the first component. Throughout the entry, |TE| signals a translation equivalent to be supplied by the user. In addition, the entry provides structured space for information pertaining to the translation equivalent: grammatical information, labels, explanations, etc.

Although the Estonian–X dictionary database (EXDD) is still being improved and updated, the system is used in-house at present for compiling three bilingual dictionaries (Estonian–Ukrainian, Estonian–Udmurt, Estonian-Finnish).

4. Dictionary management system for bilingual dictionaries

The dictionary management system for bilingual dictionaries uses the same software as the EELex dictionary administration system. The management system is based on a standard XML schema following the structure of a typical bilingual dictionary, and on a standard dictionary layout.

The user interface enables users to create their own dictionary and to adjust the system to their own needs. For every dictionary application appropriate parameters can be selected in the following four domains:

(a) selection of source and target languages accompanied by automatic keyboard switching and a spelling checker option available during the editing process;

(b) layout design: the user can decide upon the style of the elements as well as the markers of the elements or element groups;

(c) the morphological interface enables automatic generation of a morphological description of the Estonian headword: inflected forms, indexes of part of speech and inflectional paradigm; the rule-based morphological system will generate a morphological description for unknown words as well (Viks 2000);

(d) a future option enables modification of entry structure by addition, deletion or rearrangement of its elements.

Figure 3 presents an example of the window for creating a user's dictionary.

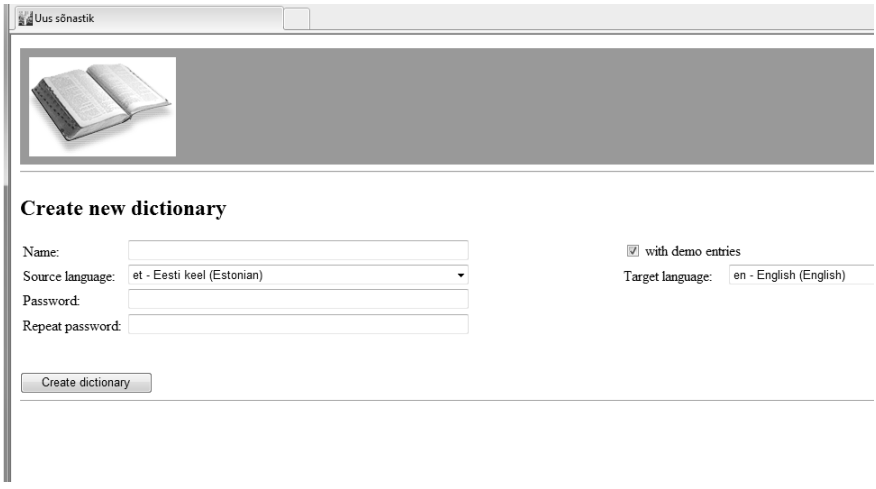


Figure 3. Dictionary creating window

A public version of the Dictionary management system for bilingual dictionaries will be released as freeware (http://exsa.eki.ee).

5. Conclusion

The dictionary management system for bilingual dictionaries enables the user, through the user interface, to design their own dictionary in a web environment, choosing the languages, modifying the entry layout and structure, and using, upon request, a ready-made description of the source language. This will economize on compilation time and improve the output quality as the resulting dictionaries represent universal re-usable language resources in a standard format.

Acknowledgements

This work was supported by the National Programme for Estonian Language Technology (2006-2010) and by the project SF0050023s09 ("Modelling intermodular phenomena in Estonian").

References

JOFFE, D., MACLEOD, M. and DE SCHRYVER, G.-M. (2008). The TshwaneLex Electronic Dictionary System. In E. Bernal and J. DeCesaris (eds). *Proceedings of the XIII EURALEX International Congress (Barcelona, 15-19 July 2008)*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra: 421-424.

LANGEMETS, M., LOOPMANN, A. and VIKS, Ü. (2006). The IEL Dictionary Management System of Estonian. In G.-M. de Schryver (ed.). *DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writing Systems*. Turin: Turin University: 11-16.

LIIV, M., MELTS, N. and ROMET, A. (eds) (1997-2009). *Eesti-vene sõnaraamat 1-5*. Tallinn: Eesti Keele Sihtasutus.

MANGEOT, M. (2006). Dictionary Building with the Jibiki Platform. In E. Corino, C. Marello and C. Onesti (eds). *Proceedings of the XII EURALEX International Congress (Turin, 6-9 September 2006)*. Turin: Turin University: 185-188.

VIKS, Ü. (2000). Tools for the Generation of Morphological Entries in Dictionaries. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis and G. Stainhaouer (eds). *Second International Conference on Language Resources and Evaluation. Proceedings*. Athens: 383-388.

Lexicography in the grid environment

Héctor Martínez[1], Marta Villegas[1],
Núria Bel, Santiago Bel, Francesca Alemany
Universitat Pompeu Fabra, Barcelona

Abstract

The research reported in this paper is part of the activities carried out within the CLARIN – Common Language Resources and Technology Infrastructure – project. CLARIN is a large-scale pan-European project to coordinate and make language resources and technology available and readily useable. CLARIN is devoted to the creation of a persistent and stable infrastructure serving the needs of the European Humanities and Social Sciences (HSS) research community. HSS researchers will be able to efficiently access distributed language resources and apply analysis and exploitation tools relevant for their research questions. In this paper we present a real case of language technology usage, addressed as a CLARIN scenario: the building of a support workflow for lexicographers who are creating a dictionary of Spanish as a second language by automatically clustering the corpus concordances with which they work.

Keywords: computational lexicography, corpus concordance clustering, grid, Hamming distance, Jaccard index, lexical similarity, orchestration, stemming, text clustering, text similarity, web services.

1. Introduction

Distributed or *grid* computing allows a great number of new technical possibilities for linguistic research, be it access to remote data or to massive computation processes, which were unavailable until recently. In a grid environment, researchers have access to Web Services, which enable them to access functionalities on a remote system. In this new scenario, researchers define their experiments as a workflow or sequence of operations where each operation invokes a Web Service.

Such architecture poses important challenges that need to be addressed: interoperability, integration between different tools and resources, workflow edition and orchestration.

2. Use case

In order to test and validate our research, we defined and implemented a real use case: providing technical support to lexicographers who are creating a dictionary of Spanish as a second language, viz. the *Diccionario de Aprendizaje del Español como Lengua Extranjera* (DAELE). The work is developed with the support of a Lexical Markup

[1] {hector.martinez,marta.villegas}@upf.edu

Framework – ISO 24613 – (Bel *et al.* 2006) compliant lexicographical platform, Coldic.[2]

Words in the core vocabulary of a language are very frequent and highly polysemous. Table 1 gives the number of occurrences in the *Corpus de Referencia del Español Actual* (CREA) and the number of senses listed in the *Diccionario de la Real Academia de la Lengua* (DRAE) for four core nouns.

Form	Occurrences	Senses
puente (‘bridge’)	6,066	15
prueba (‘test’)	15,987	14
camino (‘path’)	29,161	8
sentido (‘sense’)	50,891	11

Table 1. Corpus occurrences from the CREA and senses provided by the DRAE

Lexical entries need to provide learners of Spanish with information about usage and most frequent co-occurrence patterns. Lexicographers use lists of concordances extracted from corpora to find these patterns indicating particular meanings (Firth 1975). Often they have to struggle with a large number of occurrences for a particular word. Analyzing these large and unstructured lists becomes a highly time-consuming task. The goal, therefore, is to provide lexicographers with clustering processes that automatically structure concordances in a useful manner.

3. System architecture

The system uses lexical similarity as the measure to cluster corpus occurrences together. We calculate lexical similarity on the basis of the number of content words – non-stop words – that two occurrences have in common. The different metrics used to measure similarity are explained under point 3.

Each operational step in the system is a module that can be hosted on a different server:

1. Corpus query: The word to be analyzed is queried in the corpus and its KWIC concordances are retrieved. For syntactic and statistical reasons, only nouns have been analyzed so far using this system.
2. Stop word removal: The stop words in each occurrence are marked so that they can be discarded in further steps.

[2] Coldic: http://sourceforge.net/projects/coldic/

3. Metrics: The core of the system takes each possible pairing of occurrences A and B from the whole set of concordances and calculates a vector V_{AB}, each position of this vector being a given established metric from the following list (Rodríguez Hontoria 2003):
    a. Hamming index: Ordinal edit distance. It measures the number of lexical elements that appear in the same position in both A and B.
    b. Jaccard index: Proportion between the set intersection of A and B and the set addition between them. It describes the amount of lexical overlap between A and B, disregarding word order.
    c. Coverage index: Total representativity of a given occurrence A or B over the total bag-of-words set $A \cap B$.
    d. Equal element at a given position: It describes whether occurrences A and B share a lexical item at one relative position from their center.
4. Stemming: A brute-force stemmer is applied to the concordances. It does not use POS tagging, but a simple list of suffixes such as "-es", and "-mente". The stemmer introduces a certain number of false splits like "lugar" (En. 'place'), which is a noun, being split as "lug+ar", although "-ar" is an infinitive suffix. The noise introduction of this strategy has not been considered significant, and needing only a suffix list makes the system's linguistic portability easier than having to use full-fledged lemmatization.
5. Vector generation: Variants of the metrics are calculated using stemmed units or whole words, and providing different parameters for metrics. The system does not use any chunks larger than monograms for its calculations. For a given concordance in a corpus output set of size N, N-1 vectors of this form are calculated. All the metrics that have been chosen are symmetrical, so $V_{AB} = V_{BA}$, and concordances are not compared with themselves, which means that for any N, $N/2 - N$ vectors are calculated.
6. 1-to-1 comparison: An unsupervised K-means algorithm (Witten and Heibe 2005) clusters each of the vectors for V_{AB} with a fixed A value (noted as the V_{AX} set), from V_{A1} to V_{AN}, to determine whether they describe vectors deemed similar or different. The result of this process is that each V_{AX} set is divided in two groups: a group for the concordances that are considered similar to A by the system and a second group for those who are not.
7. Final clustering: The previous step is not enough to present a coherent list of clustered concordances, since there can be incoherent classifications (*i.e.* the system can say that F is similar to G and H, but H can be found to be different from G). This final step presents the original concordances in a series of clusters without repetitions or overlaps.

Figure 1 describes the system architecture, beginning at the first request to analyze the occurrences of a given word, to the tagging of stop words, the generation of vectors using stemming and different metrics (steps 3, 4 and 5) and the successive steps of 1-to-1 comparison and final clustering.

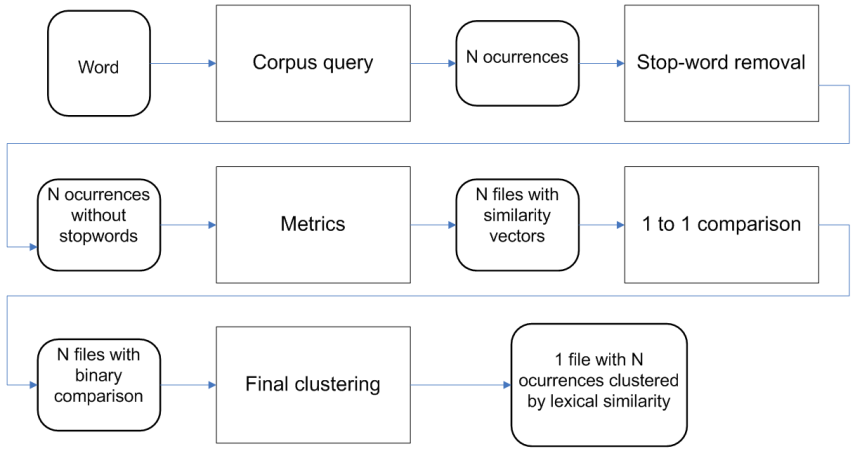


Figure 1. System architecture

In order to achieve the integration of these tools in a single process, we made them available via Web Services. Each of the boxes in Figure 1 represents a Web Service, while the round-corner boxes represent the exchanged data between modules.

entre	la	**ex**	**primera**	**dama**	filipina	y
en	La	**ex**	**primera**	**dama**	y	senadora
Claude	Pompidou	**ex**	**primera**	**dama**	de	Francia
senadora	y	**ex**	**primera**	**dama**	arrasaría	Tiene
entre	ellos	la	**primera**	**dama**	**estadounidense**	Laura
portavoz	Aunque	la	**primera**	**dama**	**estadounidense**	no
rival	de	la	**primera**	**dama**	**Hillary**	**Clinton**
enfrenta	a	la	**primera**	**dama**	**Hillary**	**Clinton**
negocios	de	la	**primera**	**dama**	**peruana**	La
negocios	de	la	**primera**	**dama**	**peruana**	BUSCADOR
Hillary	**Clinton**	una	**primera**	**dama**	tradicional	con
Hillary	**Clinton**	La	**primera**	**dama**	asegura	que
candidatura	de	la	**primera**	**dama**	En	total
candidatura	de	la	**primera**	**dama**	Es	el
Gobierno	incluida	la	**primera**	**dama**	tenían	millonarios
también	quiero	ser	**primera**	**dama**	se	manifestaron
y	actividades	como	**primera**	**dama**	La	prensa
l	Senado	La	**primera**	**dama**	no	se

Table 2. Cluster output sample

4. Results

Table 2 shows a sample of the clustering output for the word ‘dama’ (En. ‘lady’). It shows a single cluster that shares the collocation ‘primera dama’ (En. ‘first lady), although it could be divided in a more precise manner by grouping together the sub-clusters that share the elements in bold.

Table 3 shows the gold-standard evaluation for a set of words (En. ‘melanoma’, ‘uprising’, ’scrap iron’, ’hook’, ’hen’, ’drum’ and ‘lady’). ‘Unclustered’ indicates the number of occurrences that are not included in any cluster, ‘groups’ indicates the number of clusters of at least size 2 that the system generates.

	Precision	Recall	Occurrences	Unclustered	Groups
Melanoma	1,000	0,411	38	38	0
Alzamiento	1,000	0,562	41	41	0
Chatarra	1,000	0,703	71	71	0
Gancho	0,899	0,690	113	65	20
Gallina	0,904	0,685	125	74	14
Tambor	0,787	0,682	127	60	15
Dama	0,765	0,384	208	90	23

Table 3. System output evaluation

As we can see, the system does not cluster concordances for small corpus query outputs, but begins to propose clusters as the concordance numbers grow. No groups are suggested for low-frequency words like ‘melanoma’, ‘alzamiento’ and the system leaves all their occurrences unclustered.

The system also generalizes results, *i.e.* it provides coarser clusters for concordances which had been considered to be in finer clusters in the gold standard.

5. Further work

The ongoing work includes:

1. Streamlining the system to process verbs and adjectives. Prepositions and other stop words are ignored in all of the clustering metrics for nominals, since the system measures lexical similarity. Prepositional government, however, is an important aspect for verbal and adjectival classification that should be taken into account;
2. Acquiring more user feedback to determine whether the current cluster granularity is appropriate or has to be tuned to be finer or coarser;

3. Generating new stemmers for other languages (see 3.4). A similar degree of quality of the results is expected for any other Romance language;
4. Experimenting with a linear combination of the similarity metrics in order to provide a scalar – that is, a single number – for the lexical similarity between two occurrences. A scalar measure would spare the system from the need for clustering and would improve the overall efficiency, since each pairing of two concordances would have a number indicating their overall similarity, instead of having to resort to a machine learning process to determine how similar they are.

6. Conclusions

Web Services enhance the integration of different processes in long process pipelines, but in order to be fully interoperable, Web Services need more specialized typing for their service descriptors. The Web Service Description Language (WSDL) standard provides data descriptions based on simple types such as “string” or “integer”. This type system is clearly not enough, and new typing methods are required. As shown in Figure 1, we integrate different tools in a single workflow. This means that input and output must be compatible; for instance, we need the KWIC output to be the input of the vectorization process. This compatibility requires a higher service-semantics level that would provide input or output types like “KWIC concordance” as opposed to basic, blank types like “string” or “integer”.

Organizing workflows of successive Web Services with decision branching or iterative steps can be troublesome. In addition, some of the steps require large amounts of processing. This is the case of the vectorization process, where for each occurrence example the system compares it with the rest of the samples in order to generate the corresponding vectors. Slow computational processes with complex workflows need to depend on automated tools to allow the systems to run overnight. Running workflows on remote, specialized servers means that users have at their disposal enough computing resources to run the analyses and experiments they need.

Acknowledgements

The CLARIN project in Spain is co-funded by the 7FP of the EU (FP7-INFRASTRUCTURES-2007-1-212230) and the Spanish *Ministerio de Educación y Ciencia* (CAC-2007-23) and *Ministerio de Ciencia e Innovación* (ICTS-2008-11). Furthermore, the *Departament d'Innovació, Universitats i Empresa* of the *Generalitat de Catalunya* funded the development of a demonstrator that guarantees the integration of the Catalan language in CLARIN.

References

BEL, N., FRANCOPOULO, G., GEORGE, M., CALZOLARI, N., MONACHINI, M., PET, M. and SORIA, C. (2006). Lexical Markup Framework (LMF). In *Proceedings of International Conference on Language Resources and Evaluation*, Genoa.

CLARIN project web site http://www.clarin.eu.

FIRTH, J.R. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*. Oxford: Philological Society: 1-32.

LMF project web site, http://www.lexicalmarkupframework.org/.

RODRIGUEZ HONTORIA, H. (2003) Similitud semántica. In *Lexicografía computacional y semántica*. Barcelona: Edicions de la Universitat de Barcelona: 119-126.

TAYLOR, I.J., DEELMAN, E., GANNON, D.B. and SHIELDS, M. (eds) (2006). *Workflows for e-Science: Scientific Workflows for Grids*. Berlin-New York: Springer Verlag.

WITTEN, I.H. and EIBE, F. (2005). Data Mining*: Practical machine learning tools and techniques, 2nd Edition*, San Francisco: Morgan Kaufman.

The "Online Bibliography of Electronic Lexicography" (OBELEX)

Carolin Müller-Spitzer[1], Christine Möhrs[1]
Institut für Deutsche Sprache, Mannheim

Abstract

In this paper, we present the "Online Bibliography of Electronic Lexicography" (OBELEX). In addition to making general conceptual remarks, we evaluate the sources and describe the search options of OBELEX.

Keywords: bibliography, electronic lexicography, online.

1. Conceptual remarks

Digital or electronic lexicography has gained in importance in the last few years. This can be seen in the growing list of publications focusing on this field. In the OBELEX bibliography (http://www.owid.de/obelex/engl), the research contributions in this field are consolidated and are searchable by different criteria. The idea for OBELEX originated in the context of the dictionary portal OWID, which incorporates several dictionaries from the Institute for German Language (www.owid.de). OBELEX has been available online free of charge since December 2008 (*cf.* Figure 1).

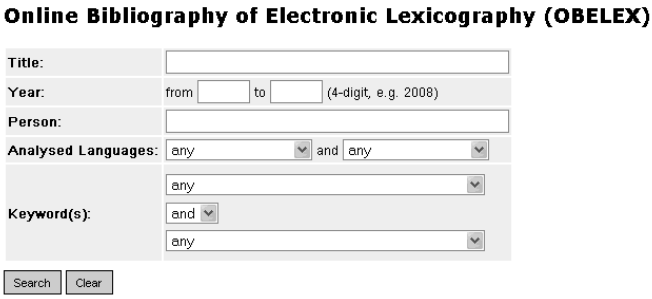


Figure 1. Online Bibliography of Electronic Lexicography.

OBELEX includes articles, monographs, anthologies and reviews published since 2000 which relate to electronic lexicography, as well as some relevant older works. Our particular focus is on works about online lexicography. Information on dictionaries is currently not included in OBELEX. However, we are working on a

[1] {mueller-spitzer,moehrs}@ids-mannheim.de

database which contains information on online dictionaries as a supplement to OBELEX. All entries of OBELEX are stored in a database. Thus, all parts of the bibliographic entry (such as person, title, publication or year) are searchable. Furthermore, all publications are associated with our keyword list; therefore, a thematic search is also possible. The subject language is also noted. For example, it is possible to search for all metalexicographic works from the field of "bilingual/multilingual lexicography" that deal with "English" and "German" (from a metalexicographic point of view).

With this type of content, the OBELEX bibliography supplements in a useful way other bibliographic projects such as the printed "Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung" by H.E. Wiegand (2006/2007), the "Bibliography of Lexicography" by R.R.K. Hartmann (2007), and the "International Bibliography of Lexicography" of Euralex (*cf.* also DeCesaris and Bernal 2006). OBELEX differs from all these bibliographic projects by its strong focus on electronic lexicography and its ability to retrieve bibliographic information:

- The Wiegand bibliography is surely the most extensive, but it does not focus on electronic lexicography and it is to be discontinued. Furthermore, searching for a specific publication is not easy, since the forthcoming register volume has not yet been published (*cf.* Dziemianko 2008). In addition to this, the main focus is on dictionary research within the field of German Studies. In contrast, OBELEX includes all international research (but only with respect to digital lexicography).
- The Euralex bibliography is an "International Bibliography of Lexicography" made "for lexicographers, by lexicographers") (*cf.* the website http://euralex.pbworks.com/). It comprises a great amount of bibliographic information about articles and books; new references are added by means of a Wiki system. There are plans in place for the Euralex bibliography to include all publications from the Euralex conference proceedings in the future. (The current online version does not include all articles from Euralex conference proceedings.) However, other periodicals or journals are not systematically included. By contrast, OBELEX systematically lists all relevant publications, such as contributions from the *International Journal of Lexicography*, the journal *Dictionaries* or the Euralex proceedings, but only the contributions related to electronic lexicography. Therefore, if a researcher is interested in *all* publications in the Euralex proceedings, the Euralex bibliography is the right choice. If s/he wishes to search in the field of electronic lexicography, the use of OBELEX is advisable.
- The Hartmann bibliography is extensive and international. However, for anthologies, only the title of the book is listed, not the individual articles. Therefore, it is not possible to search for specific articles or reviews in this bibliography.

2. Evaluation of sources

As mentioned above, in OBELEX, sources have been evaluated systematically since 2000. This limit was set for pragmatic reasons. However, there is very relevant literature about electronic lexicography from before 2000 which a user of OBELEX might miss. Therefore, often cited articles or books from the 1980s or 1990s are also included in OBELEX, for instance, the relevant volumes of the Handbooks of Linguistics and Communications Science ("Handbücher zur Sprach- und Kommunikationswissenschaft"). However, the evaluation of sources before 2000 is not comprehensive.

In addition to the systematically evaluated sources, further relevant literature is included in OBELEX. These are mainly monographs from the field of electronic lexicography and articles from journals besides the ones named above. Reviews are also integrated because they often include interesting metalexicographic elements relating to critical evaluation of electronic dictionaries and are quite often not easily accessible. As far as possible, abstracts are included in OBELEX, especially for articles from conference proceedings. These serve as a first insight into the article and may help the user to find appropriate literature. In the future, current issues of the sources will be examined systematically in order to continuously enlarge OBELEX.

The systematically evaluated literature (only contributions to electronic lexicography) in OBELEX includes the following sources:

- Dictionaries: Journal of the Dictionary Society of North America
- Hermes
- International Journal of Lexicography
- Lexikos: Annual Journal of the AFRILEX
- Proceedings of the EURALEX conferences
- Proceedings of the "International Symposium on Lexicography"
- Lexicographica (International Annual and series maior)
- Handbooks of Linguistics and Communications Science (Handbücher zur Sprach- und Kommunikationswissenschaft – HSK) vol. 5.1-5.3

3. Search options

There are different fields for searching in OBELEX. Firstly, there is the title box. By entering a word, a full-text search is carried out on all titles listed in OBELEX. "Title" always means the title of the articles or monographs, not of the series, journals, or anthologies.

3.1. Search by title

If a search is conducted for titles that include the word “vision”, the result illustrated in Figure 2 is given. This demonstrates that the search is a real full-text search because all results except the title Hahn/Klosa/Müller-Spitzer/Schnörch/Storjohann (2008) would have been expected. In case of reviews, after the title the word “(Review)” is always inserted. Thus, it is possible to search by title in order to find all the reviews which are stored in OBELEX.

Search Results

The listed order can be changed by clicking on "Person" or "Year".

	Person	Year ▾	
☐	Tarp	2009	Beyond Lexicography: New Visions and Challenges in the Information Age
☐	Meijssen	2009	The Philosophy behind OmegaWiki and the Visions for the Future
☐	Smit	2008	Henrik Gottlieb and Jens Eric Mogensen (Editors): Dictionary Visions, Research and Practice. Selected Papers from the 12th International Symposium on Lexicography, Copenhagen 2004 (Review)
☐	Fuertes-Olivera	2008	Henrik Gottlieb, Jens Erik Mogensen (Editors): Dictionary Visions, Research and Practice. Amsterdam/Philadelphia: John Benjamins Publishing Co. 2007 (Terminology and Lexicography Research and Practice 10) (Review)
☐	Hahn/Klosa/Müller-Spitzer/Schnörch/ Storjohann	2008	elexiko - das elektronische, lexikografisch-lexikologische korpusbasierte Wortschatzinformationssystem. Zur Neukonzeption, Erweiterung und Revision einzelner Angabebereiche
☐	Gottlieb/Mogensen	2007	Dictionary Visions, Research and Practice. Selected Papers from the 12th International Symposium on Lexicography, University of Copenhagen, April 29th - May 1st, 2004

Figure 2. Search Results

3.2. My bibliography

For every search result, it is possible to include selected titles in “My bibliography”. This personal bibliography is accessible for 365 days from the user’s computer (over cookies) and can also be presented in a view customized for printing. An export of “My bibliography” in BibTex format is also provided.

3.3. Search by person

The search in the search field “Person” is an incremental search. Thus, by typing in the first letters of an author’s name such as “de”, all appropriate people included in OBELEX appear and can be selected from a list.

3.4. Search by publication year

The search for publications in OBELEX can also be delimited by publication year. For instance, it is possible to search for all titles on the subject of electronic lexicography published before 2000. For this particular search, the “from”-field should be left empty and “1999” should be entered in the “to”-field.

3.5. Search by keyword or analysed language

Two of the most important functions of OBELEX are the options to search by keyword and by analysed language (*cf.* Figure 3). These fields allow a thematic search

for publications. The keyword list integrated in OBELEX originated from a comparison of different glossaries, registers, indices, or subject catalogues. From a pragmatic point of view, it was important to configure the keyword list in a not too complex and relatively general way because the tagging is made only by reading the abstracts. The whole article is consulted in just a few cases (where there is uncertainty about which keyword to assign). The language field does not apply to the language the article is written in but to the topics discussed. For example, if an article is published in English about Spanish electronic lexicography, only "Spanish" is recorded as a language in the database. Therefore, these fields allow searches such as all titles concerning the "Italian" language and the topic "online lexicography".

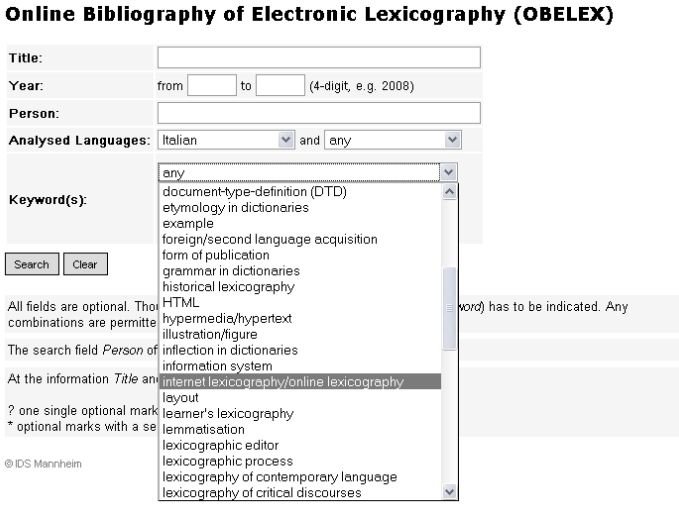


Figure 3. Search for 'Analysed Language' + 'Keyword'

4. Perspectives

With OBELEX, we hope to provide an extensive service for all researchers working in digital lexicography and dictionary research. For the future, the main task in OBELEX is to enlarge the bibliography. Therefore, we appreciate any suggestions for publications which may be included in OBELEX. As said before, besides this, we work on a supplement to OBELEX, *i.e.* a database of online dictionaries. In this database, currently about 800 online dictionaries are labelled by more than 40 properties, such as "including an online tutorial: yes/no", "pronunciation as audio files: yes/no" etc. We plan to publish this database on the OBELEX site during 2010.

References

DECESARIS, J. and BERNAL, E. (2006). Lexicography as Reported: the EURALEX Conference Proceedings Bibliography (1983-2004). In E. Corino, C. Marello and C. Onesti

(eds). *Proceedings of the Twelfth EURALEX International Congress*, Torino, Italia, September 6th-9th, 2006. Alessandria: Edizioni dell'Orso: 1241-1247.

DZIEMIANKO, A. (2008). Review: Wiegand, Herbert Ernst. Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung: Mit Berücksichtigung anglistischer, nordistischer, romanistischer, slavistischer und weiterer metalexikographischer Forschungen. *International Journal of Lexicography*, 21(3): 359-360.

EURALEX: *International Bibliography of Lexicography*. June 10, 2009 <http://euralex.pbworks.com>.

HARTMANN, R.R.K. (2007). *Bibliography of Lexicography*. June 10, 2009, <http://euralex.pbworks.com>.

HAUSMANN, F.J., REICHMANN, O., WIEGAND, H.E., and ZGUSTA, L. (eds) (1989-1991). *Wörterbücher. Dictionaries. Dictionnaires: Ein internationales Handbuch zur Lexikographie. An International Encyclopedia of Lexicography. Encyclopédie internationale de lexicographie (HSK 5.1-5.3)*. Berlin/New York: de Gruyter. (Handbücher zur Sprach- und Kommunikationswissenschaft. Handbooks of Linguistics and Communication Science. Manuels de linguistique et des sciences de communication 5.1.-5.3).

WIEGAND, H.E. (2006/2007). *Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung: mit Berücksichtigung anglistischer, nordistischer, romanistischer, slavistischer und weiterer metalexikographischer Forschungen*. Berlin / New York: de Gruyter.

From lexical database to intelligent vocabulary trainers

Cornelia Tschichold[1]
Swansea University

Abstract

Computer-assisted language learning needs better lexical databases in order to produce better software for vocabulary learning. This paper attempts to give some guidelines for the construction of a dedicated lexical database for vocabulary learning purposes.

Keywords: vocabulary, CALL, lexical databases, second language learning, polysemy.

1. Background

The field of computer-assisted language learning (CALL) has produced vocabulary trainers for some time now, and many language learners feel they need to improve their vocabulary knowledge, so the incentive is clearly there to produce good software that offers learners help with this task. What we find on the market, however, tends to be drill-and-kill-style vocabulary trainers, or software to produce the electronic equivalent of filing card systems, or – in the best of cases – a reasonably useful combination of these. A fundamental problem with all such material is the rather simplistic view of vocabulary knowledge promoted by these tools. L2 words are either (typically) given a single L1 translation to be memorized, or (less frequently) a dictionary-style entry is used, with a whole series of translational equivalents. For absolute beginners, the single translational equivalent is probably appropriate, but once such a first rudimentary structure of "L2 word = L1 word" is in place in the learner's mental lexicon, more sophisticated lexical input is needed, without however going all the way to overwhelm the learner with a complete dictionary entry. In the following, a dedicated lexical database is proposed which can form the basis of intelligent e-learning material. The field of lexicography, and learner's lexicography in particular, obviously has something to offer to such an enterprise.

One of the most robust findings on vocabulary acquisition (Nation 2001) is that individual vocabulary items need to be repeated several times before a sufficiently rich mental representation can be constructed by the learner. The number of repetitions needed depends on a multitude of factors, *e.g.* cognateness and a whole set of word-inherent qualities. Verbs tend to be harder to learn than nouns, for example. Very polysemous words also have a higher learning burden than monosemous words. Learners'

[1] Department of English Language and Literature, Swansea University, c.tschichold@swan.ac.uk

assumptions of polysemy in L2 words will in part depend on their L1, which might have similar polysemy for individual words. Divergent polysemy, however, (see examples a to c) is likely to lead to a particularly high learning burden.

L1 German *Schatten* > L2 English *shade / shadow*

L1 English *river* > L2 French *fleuve / rivière*

L1 French *bureau* > L2 English *office / desk*

Many highly frequent words, especially verbs, also have a high learning burden due to the sheer number of their possible subsenses. While a single translation for the English verb *run* is appropriate for complete beginners, very soon other translations will be needed in context, *e.g.* for the expression *to run a shop*. The entry for *run* even in a learner's dictionary can easily "run" to several pages. Faced with such a depth of vocabulary knowledge to be learnt, learners need to be presented with a well-structured sequence of word forms and usages to allow for efficient uptake. Such a principled approach to teaching should probably best start with the prototypical sense for each word (not necessarily the most frequent sense), then proceed in order of difficulty, before taking in collocations and other phraseological uses of the word.

Giving learners access to corpus examples has been proposed as a way of presenting the target word in many different contexts,[2] but this method risks complicating the learner's task too much when the concordance lines contain examples where the target word is used in a different sense. When a learner is trying to learn the noun *pupil* for example, concordance lines are likely not only to include the 'student' sense, which is the one our learner is probably interested in, but also the 'part of the eye' sense. This effect is not a welcome one for beginning learners, who are still trying to master the most frequent words of their foreign language. Frequent words by their very nature are also those words where polysemy is particularly prevalent. Adding random corpus examples to the learning material will therefore not solve the problem of finding a number of suitable context sentences for the target vocabulary.

2. Lexical databases for learners

To make more sophisticated vocabulary trainers possible and be prepared for porting them to ever more mobile devices suitable for learning any time anywhere, we need lexical databases that contain more information than present ones. In addition to data found in dictionaries, the database needs to contain information about word frequency by subsenses. Modern learner dictionaries have already begun to move in this direction by reducing the number of subsenses per headword, and by adding frequency information. For the verb *run*, this reduction can amount to as much as going from several dozen to under twenty subsenses. What we do not yet have is frequency information per subsense rather than just for the headword.

[2] See "Gerry's Vocabulary Teacher" <www.cpr4esl.com/gerrys_vocab_teacher> for a good example of this approach.

A lexical database for vocabulary learning should include several suitable example sentences illustrating each subsense. This requires linking the lexical database to corpus examples where the words have been disambiguated for sense, a task which can be speeded up with NLP tools, but which cannot be done reliably entirely automatically. Part of the problem can be traced back to the techniques used in corpus linguistics (Gardner 2007), which do not lend themselves easily to applications in CALL. While the construction of such a database with corpus examples might seem to be a dauntingly huge task, we should remember that we are dealing with learner language of a few thousand lexemes at most. The main task in the construction of a dedicated lexical database for vocabulary learning lies in the selection of what to present to the learner and in what order, along with the collection of suitable examples for several stages of learning. Because of the need for repetition, a single example or definition is not enough; even one example for each subsense is likely to be insufficient. Individual word entries have to be rich and varied enough to allow for generous amounts of learning material.

Such a database could be seen to have some parallels with learner's dictionaries or bilingualized dictionaries, but while it should not be confused with a dictionary, much of the information needed for the database can be found in dictionaries. Most of the syntactic and semantic information needed can be extracted from learner's dictionaries, and possibly some examples as well.

Deciding on the appropriate subsenses for polysemous words (and their order) will be one of the first difficult decisions to make for each word. Subsenses can be found in dictionaries, both monolingual and bilingual, and in more specialized databases such as valency dictionaries (*e.g.* Herbst *et al.* 2004), WordNet (Miller *et al.* 1990) and in West's (1953) General Service List.[3] The problem is the divergence among these sources. To borrow an example from Atkins & Rundell (2008: 154), who use the verb *argue* to illustrate the identification of subsenses from corpus evidence and come up with four "lexical units" – illustrated with corpus examples and linguistic features –, for a teaching-oriented lexical database, a choice would have to be made about which of the subsenses of *argue* to include, and in which order. West lists two of the subsenses as reasonably frequent, so a first version could assume two subsenses for *argue*. The next step is to establish the order of the subsenses, especially which of them will be taught first and thus provide a kind of prototypical meaning to the language learner. The decision on the first meaning to teach should consider frequency, prototypicality and possibly also the wider context of the word family (Bauer and Nation 1993). All subsenses and their syntactic contexts will then need to be illustrated with a series of examples, graded by difficulty. In the example shown in Table 1, difficulty of context is expressed using the frequency bands of the words around the target word, *i.e.* a "4K" context means that the words in the sentence come from the 4000 most frequent words in English.

[3] West's list is now very dated, but it does have the considerable advantage of not just listing subsenses, but also giving frequencies for each of them, along with an example.

3. An example

The database entry for the words to be learned will have the type of information typically found in dictionaries, *i.e.* pronunciation, inflected forms, syntactic information, and meanings, along with translational equivalents if required. In addition, the database will contain many examples of usage, coded for the form and meaning of the target word and for the difficulty of the context. The database entry for the noun *model*, for example, might contain the information shown in Table 1 linked to an ordered sequence of example sentences.

No	Context	Subsense	Sg/Pl	Example
1	1K	A [small version]	sg	From the distance the church looked smaller than I had expected, like a **model** of the real building.
2	2K	A	pl	The children make the **models**, including sheep, dogs and horses, for gift shops to sell.
3	1K	B [type]	sg	The newer **model** with twice as much memory costs £385.
4	4K	B	pl	A study reveals that car producers made little gains in fuel economy ratings for their new car **models.**
5	4K	C [role model]	pl	The monastery and the military provided **models** for the early schools.
⋮	⋮	⋮		⋮
9	3K	D [fashion model]	sg	Win one of our special beauty make-over sessions and you will feel like a top **model**.
⋮	⋮	⋮		⋮
14	4K	E [description]	pl	This chapter begins by examining several analytical **models** and looks at their usefulness as tools for analysis.
⋮	⋮	⋮		⋮
…	…	…	…	…

Table 1. Database entry for model

Information about the frequency of the different forms of a lemma (singular and plural forms in the case of *model*) should also be included and taken into account when selecting material to include. Should the plural form of a noun occur considerably more frequently than the singular form, then the database ought to reflect this fact, especially if the different inflectional forms correlate strongly with a particular subsense (Stubbs 2001).

Drawing on such a database, a CALL system could present learners with many different example sentences as they work their way through the necessary number of repetitions for each of their target words. This would enable learners to systematically

enlarge their vocabulary in a much more efficient way. A dedicated vocabulary CALL programme could start the learning process with the help of a translational equivalent at the first encounter. Once this first link has been established, knowledge about the new word needs to be consolidated through repeated encounters of the target word in different inflectional forms and syntactic contexts, gradually adding further subsenses, in order to progressively arrive at a rich mental representation of the target word.

4. Conclusion

Intelligent vocabulary CALL needs to integrate the advantages that incidental vocabulary acquisition via extensive reading provides into the material for intentional vocabulary study. A dedicated database as sketched here could provide the raw material for CALL exercises that combine the best of intentional and incidental vocabulary e-learning.

References

ATKINS, S. and RUNDELL, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

BAUER, L. and NATION, I.S.P. (1993). Word families. *International Journal of Lexicography*, 6: 253-279.

GARDNER, D. (2007). Validating the Construct of *Word* in Applied Corpus-based Vocabulary Research: A Critical Survey. *Applied Linguistics*, 27/2: 241-265.

HERBST, T., HEATH, D., ROE I.F. and GÖTZ, D. (2004). *A Valency Dictionary of English*. Berlin: Mouton de Gruyter.

MILLER, G., BECKWITH, R., FELLBAUM, C., GROSS, D. and MILLER, K. (1990). Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, 3/4: 235-244.

NATION, I.S.P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

STUBBS, M. (2001). *Words and Phrases.* Oxford: Blackwell.

WEST, M. (1953). *A General Service List of English Words*. London: Longman.

Multiple access routes

The dictionary of Bavarian dialects in Austria / Wörterbuch der bairischen Mundarten in Österreich (WBÖ)

Eveline Wandl-Vogt[1]
Institut für Österreichische Dialekt- und Namenlexika,
Österreichische Akademie der Wissenschaften

Abstract

In this article the Academic Austrian dialect dictionary, the *Wörterbuch der bairischen Mundarten in Österreich* (*WBÖ / Dictionary of Bavarian dialects in Austria*) is introduced. In 1993, the project *Datenbank der bairischen Mundarten* (*DBÖ / Database of Bavarian dialects in Austria*) started, with a view to the digitisation of the dictionary and additional background information. In 1998 a rationalisation concept was issued planning to complete the dictionary in 2020 as a (virtual) unit consisting of the printed dictionary and the complementary database. The *Database of Bavarian dialects in Austria electronically mapped (dbo@ema)* combines an online dictionary and a source material database. It demonstrates how visual, geo-referenced access structures and so called 'topographic navigation' can increase usability and lead to improved interdisciplinary insight.

Keywords: dialect dictionary, dialect Database, navigation, dbo@ema, geo-referenced dialect data, mapping, WebGIS.

1. Dictionary and database: the conception of a virtual unit

Dialect dictionaries are long term projects, usually not very open to modernisation and changes due to the fact that every change of aims or methods in long term projects usually yields to the investment of a lot of money.

The *Dictionary of Bavarian Dialects in Austria* (*WBÖ*) was first published in 1963 by today's Institute of Lexicography of Austrian Dialects and Names (Institut für Österreichische Dialekt- und Namenlexika). The working group established a database in 1993 to store the dictionaries base material.[2] At the moment, nearly two thirds of the base material, about 5 million mostly hand-written paper slips, are fully digitized (A sample of paper slips can be seen at http://www.wboe.at/en/hauptkatalog.aspx). In 2010 sample entries will be accessible in a web-based, interactive and geo-referenced format for the first time (*cf.* dbo@ema: http://wboe.oeaw.ac.at).

[1] Institut für Österreichische Dialekt- und Namenlexika (Institute of Lexicography of Austrian Dialects and Names), eveline.wandl-vogt@oeaw.ac.at

[2] For further information about e-lexicography at the DINAMLEX *cf.* Wandl-Vogt (2008b).

In 1998 a rationalisation concept (Straffungskonzept 1998) was issued to the *WBÖ* with the objective to complete the dictionary in 2020 as a (virtual) unit consisting of the printed dictionary on the one hand and the complementary database on the other. At the same time, both, the mediostructure and microstructure of the dictionary were adjusted and improved (*cf.* Wandl-Vogt 2004). New types of entries were established (so called Datenbankartikel [database entry]):

(1) Simple database entry (historical base material):

WBÖ 5,27: †Diaun, Gerichtsbote obVintschg. (16.Jh.), s. DBÖ

(2) Simple database entry (recent base material):

WBÖ 5,512: Trikó, M., N., Trikot, elast. Stoff; best. eng anliegendes Kleidungsstück ugs., s. DBÖ

2. Traditional access structures: the dictionary's macrostructure in a nutshell

Due to the macrostructure of the *WBÖ*, the etymological-historical headword and the highly sophisticated structure of the entry itself[3], access to the *WBÖ* is neither very functional nor user-friendly. Examples 3-5 illustrate some of the difficulties we encountered.

(3) The standard German equivalent *Apfelbaum* ('apple tree') corresponds with the *WBÖ*-headwords (*Apfel*)*pāum* and (*Epfel*)*pāum* which themselves are subentries[4] of the *WBÖ*-main-entry *Pāum* (standard German ('Baum', 'tree'; WBÖ 2,621).

(4) The etymological-historical *WBÖ*-headword *teütsch* 'German' corresponds with the standard German equivalent *deutsch* (WBÖ 5,23).

(5) The etymological-historical *WBÖ*-headwords *Tscharda, Tscharde, Tschardere* ('old house'; Hungarian; WBÖ 5,731), *Tscherper, Tschirper* ('imbecile person, old man, frayed edge tool'; Slovene; WBÖ 5,753) or *tschinkwe* ('inferior'; Italian; WBÖ 5,767) lack any standard German equivalent.

Due to this very specific headword-tradition, digitizing the *WBÖ* and linking its content with other dictionaries and databases requires a lot of effort, technical as well as lexicographical.

3. New access structures: topographic navigation

Topographic information plays an important role in dialect dictionaries (*cf.* Kühn 1982: 704f.). Within the project *Database of Bavarian Dialects in Austria*

[3] Different types of entries and article-structure *cf. WBÖ*-Beiheft 2: 14-17.
[4] (*Epfel*)*pāum* itself is a subentry of (*Apfel*)*pāum*.

electronically mapped (*dbo@ema*; 11.2007-11.2009) multiple access routes have been developed, aiming to meet user needs.[5] In the following, topographic navigation will be focused on.[6]

The aim of the newly developed system *dbo@ema* is not just to visualize database content but to enable the navigation from the map to the database and to analyse data based on a GIS. Figure 1 presents a screenshot of the visualization. It demonstrates the zooming into a certain area (*Lungau*), clicking on a certain locality (*Zederhaus*) and getting information from a linked, geo-referenced bibliography. The chosen base map is a satellite map, but might as well be any[7] *e.g.* a historical map. Data in the pop up are linkable. The user can navigate straight from the map into the data base content and vice versa (*cf.* Figure 2).

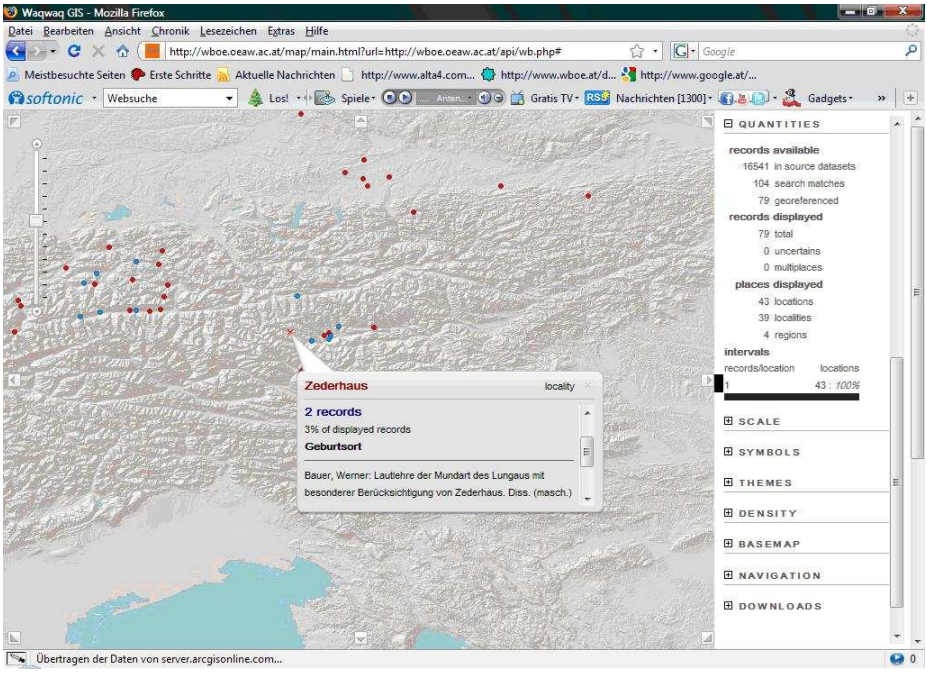


Figure 1: GIS-Application: Visualization of spatial data with ArcGIS (example: screenshot developed within dbo@ema by Vlad Atanasiu, Sebastian Arming & Eveline Wandl-Vogt: system under construction).

[5] Scholz *et al.* (2008), Wandl-Vogt *et al.* (2008a); wboe.oeaw.ac.at.

[6] To get more information about projects dealing with the mapping of linguistic corpora *cf.* Perea (2004; Catalan) and Wandl-Vogt (2008b; *DBÖ*).

[7] Realized as far as *dbo@ema* is concerned (12.2009): shaded relief, physical world, satellite map, road map, none.

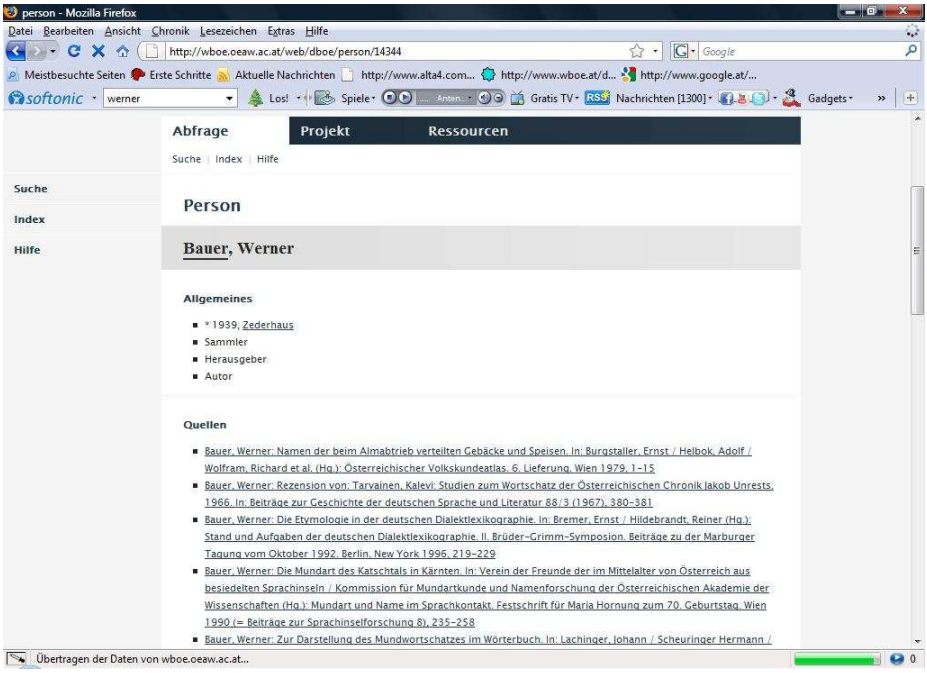


Figure 2: Website: Database content[8]

4. Conclusion and future prospects

Navigation in dialect dictionaries is often difficult and not very user-friendly. Dialect dictionaries deal with geographic information, so the access to them is based on something like "geographically based interest". Geo-referencing of corpus data provides the means for visualization (mapping) of data and meets user needs. In addition, it allows cross-linking of data from different sources and leads to improved insight and knowledge.

References

DBÖ = Institut für Österreichische Dialekt- und Namenlexika (I DINAMLEX) (ed.) (1993-). *Datenbank der bairischen Mundarten in Österreich (DBÖ).* Wien.

KÜHN, P. (1982). Typen lexikographischer Ergebnisdarstellung. In W. Besch *et al.* (eds). *Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung.* 1 Halbband. Berlin, New York: Walter de Gruyter: 702-723.

PEREA, M.-P. (2004). New Techniques and Old Corpora: La felxió verbal en es dialects catalans (Alcover-Moll, 1929-1932). Systematisation and Mapping of a Morphological

[8] Screenshot developed within dbo@ema by Thomas Heymann, Sebastian Arming and Eveline Wandl-Vogt; system under construction.

Corpus. *Dialectologia et Geolinguistica (DiG). Journal of the International Society for Dialectology and Geolinguistics*, 12: 25-45.

SCHOLZ, J., BARTELME, N., FLIEDL, G., HASSLER, M., MAYR, H. C., NICKEL, J., VÖHRINGER, J. and WANDL-VOGT, E. (2008). Mapping languages – Erfahrungen aus dem Projekt dbo@ema. In J. Strobl, T. Blaschke and G. Griesebner (eds). *Angewandte Geoinformatik 2008: Beiträge zum 20. AGIT-Symposium*. Heidelberg: Wichmann: 822-827.

Straffungskonzept (1998) = Institut für Österreichische Dialekt- und Namenlexika (1998). *Neues Straffungskonzept für das Wörterbuch der bairischen Mundarten in Österreich (WBÖ).* Wien: Masch.schriftl. – printed Version: WBÖ-Beiheft 2: 11-13; online Version: <accessed: http://www.oeaw.ac.at/dinamlex/Straffungskonzept_1998.pdf>.

WANDL-VOGT, E. (2004). Verweisstrukturen in einem datenbankgestützen Dialektwörterbuch am Beispiel des Wörterbuchs der bairischen Mundaren in Österreich (WBÖ). In S. Gaisbauer and H. Scheuringer (eds). *Linzerschnitten. Beiträge zur 8. Bayerisch-österreichischen Dialektologentagung, zugleich 3. Arbeitstagung zu Sprache und Dialekt in Oberösterreich, in Linz, September 2001*. Linz: Adalbert-Stifter-Institut des Landes Oberösterreich: 423-435.

WANDL-VOGT, E. *et al*. (2008a). Database of Bavarian Dialects (DBÖ) electronically mapped (dbo@ema). A system for archiving, maintaining and field mapping of heterogenous dialect data. In E. Bernal and J. DeCesaris (eds). *Proceedings of the XIII EURALEX International Conference* (Barcelona, 15-19 July 2008). Barcelona: Institut Universitari de Lingüistica Aplicada / Universitat Pompeu Fabra. CD.

WANDL-VOGT, E. (2008b). Wie man ein Jahrhundertprojekt zeitgemäß hält: Datenbankgestützte Dialektlexikografie am Institut für Österreichische Dialekt- und Namenlexika (I DINAMLEX) (mit 10 Abbildungen). In P. Ernst (ed.). *Bausteine zur Wissenschaftsgeschichte von Dialektologie / Germanistischer Sprachwissenschaft im 19. und 20*. Jahrhundert. Wien: Präsens: 93-112.

WBÖ = Institut für Österreichische Dialekt- und Namenlexika (I DINAMLEX) (ed.) (1963-). *Wörterbuch der bairischen Mundarten in Österreich (WBÖ).* Wien: Verlag der Österreichischen Akademie der Wissenschaften.

WBÖ-Beiheft = Institut für Österreichische Dialekt- und Namenlexika (I DINAMLEX) (ed.) (2005). *Wörterbuch der bairischen Mundarten in Österreich (WBÖ). Beiheft Nr. 2. Erläuterungen zum Wörterbuch.* Wien: Verlag der Österreichischen Akademie der Wissenschaften.

NLP tools for lexicographic applications in Modern Greek

Christos Tsalidis, Mavina Pantazara,
Panagiotis Minos, Elena Mantzari
Neurolingo

Abstract

In this paper, we introduce the language tools developed by Neurolingo LP in order to support lexicographic NLP applications and related computer-assisted activities. First, we present three infrastructure tools used for encoding morphological, semantic, and syntactic information. Second, we present some application tools such as the proofing tools for Modern Greek, *i.e.* spelling checker, hyphenator, and thesaurus. Finally, we sketch some avenues for future research.

Keywords: NLP tools, lexicographic applications, Modern Greek.

1. Introduction

In this paper, we present the R&D activities carried out by Neurolingo LP regarding the development of language tools and language resources for Modern Greek – henceforth M. Greek. Language tools can be grouped into: a) infrastructure tools: software systems for the development of language resources, *i.e.* lexicographical databases, corpus management systems, rule-writing or machine-learning workbenches, etc., and b) application tools: software components or systems built upon the language resources and utilized by end-users to access information, *i.e.* lexicon browsers and search engines, or to perform automatic text or speech processing, *i.e.* text-to-speech converters, spelling/grammar/style checkers, summarizers, machine translation systems, etc. In section 1, we describe three infrastructure tools for the encoding of different types of lexical information – morphological, semantic, and syntactic. In section 2, we present some application tools such as the proofing tools for M. Greek, *i.e.* spelling checker, hyphenator, and thesaurus. Finally, we sketch some future directions.

2. Lexical information encoding tools

2.1. Morphological information

The LEXEDIT is a software program for the compilation of computational morphological lexica. The system supports the encoding of morphosyntactic attributes – POS, gender, number, case, voice, tense, mood, person, etc. –, stylistic attributes – formal, informal, oral, slang, etc. –, domain attributes – biology, medicine, law, music, computer, etc. –, morphemic compounding – stem, prefix, infix, suffix –, syllabification, inflection and stressing rules for each lexeme.

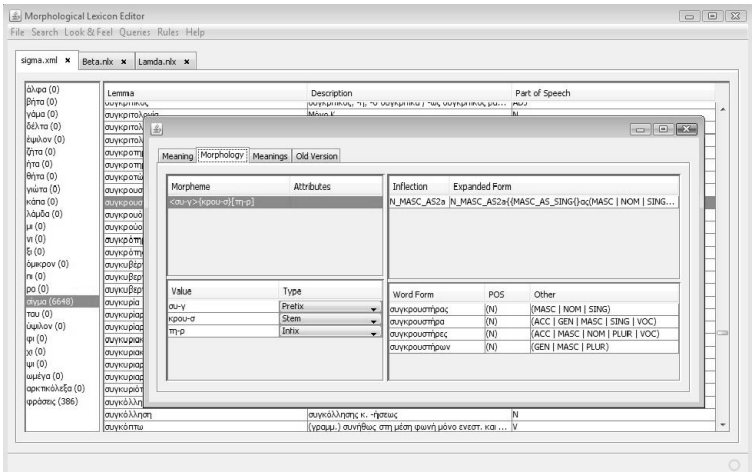


Figure 1. Screenshot of the LexEdit

The LEXEDIT is based on standards such as XML and Unicode and supports practically any language. For the description of the morphological particularities of the M. Greek language, viz. high inflection, stressing rules, morphological, phonological or graphical variants, 306 inflection rules were constructed – 135 for nouns, 96 for verbs, 50 for adjectives, 17 for participles – using sets of 191 suffix rules combined with 19 stressing rules. The LEXEDIT was used for the creation of the following lexical resources:

a) a general language morphological lexicon, which contains ~90,000 words, *i.e.* ~1,100,000 word forms with orthography, syllabification, morphological, morphosyntactic and morphostylistic information;

b) an electronic dictionary of geographic names and toponyms of Greece with ~10,000 lemmas;

c) an electronic dictionary of biomedical terms with ~10,000 lemmas.

A series of application tools such as a M. Greek speller, a hyphenator and a lemmatizer were built upon the aforementioned lexical resources.

2.2. Semantic information

The LEXICOGRAPHOS is a language independent software for editing and authoring monolingual and bilingual dictionaries and for publishing dictionaries to hardcopy, electronic or online media. It can process XML files through a user-friendly interface, which further allows integrity cross-referencing checking between different entries of the dictionary, automatic up-dating of sense numbers, full or partial entry previewing, grouping of lemmas using filtering, searching on the headwords list and/or other fields of an entry, display of hyperlinked related lemmas, selection of the information to be viewed and/or exported, spell-checking of the data.

Moreover, the system is designed to support the production of a printed PDF version of the dictionary keeping the formatting annotations through XSLT (Extensible Stylesheet Language Transformations). So far, it has been used for the editing of two hardcopy dictionaries: the "Synonyms and antonyms Thesaurus of Modern Greek" (Patakis Publications, Athens 2005) and the "Dictionary of Modern Greek as a foreign language" (University of Athens 2007).[1]

More specifically, the Thesaurus of M. Greek contains ~22,000 entries. Each entry is represented by the headword lemma[2], which is further accompanied by stylistic and domain information, *e.g.* the verb *αγκαζάρω* is informal, the noun *αιμοσφαιρίνη* is a term of Biology. Each meaning of a lemma is described through a "synset", *i.e.* a set of synonyms and/or antonyms, and may also contain example uses, when needed.

2.3. Syntactic information

The KANON – from the Greek word *κανών* "rule" – is a feature-based grammar formalism, which is used for the recognition of specific morphosyntactic patterns in the input text documents. The format of the rules resembles context sensitive extended Backus-Naur Form (BNF) rules, where every symbol is presented as a set of feature value pairs. The grammar is strongly "typed" in the sense that every feature must be previously defined together with a type which specifies the values of its instances in the rules. The rules definition incorporates the functionality of the lemmatizer and can use lexical features such as full lemma (*e.g.* the verb *to increase i.e. increase, increases, increased, increasing*), word form (*e.g.* the word form *increasing*), morphosyntactic attributes (*e.g.* noun_sing_nom, verb_pass_pres), morphological attributes (*e.g.* words ending in *-ing*), orthographic attributes (*e.g.* words starting with capital letter). The rules can be applied in a consecutive and aggregative manner. 'Consecutive' means that rules are applied in the same sequence of annotated text spans repeatedly. In other words, as long as we can apply rules and the size of the text span's sequence is decreased, the processing continues. This scheme together with the

[1] http://www.museduc.gr/docs/gymnasio/Dictionary.pdf

[2] The headword lemma is the canonical form of a lexeme, *i.e.* the singular, nominative form for nouns, the 1st person, singular, present, indicative, active form for verbs.

context free nature of rules simulates classical LR parsing[3] avoiding cycles or endless reduction loops. 'Aggregative' means that a set of rules can be applied after another set of rules. We can have as many levels of parsing as is necessary for handling different instances of "context sensitive" syntactic cases of natural languages. This formalism constitutes the core component of a number of NLP applications, such as multi-word term identification in the biomedical domain, Named Entities Recognition in the framework of text mining and information extraction, and grammar checking.

3. Application tools

3.1. Proofing tools

The proofing tools are functional components of almost all contemporary word processors and professional desktop publishing systems. They assist users – typists, typesetters, writers, translators, editors, etc. – to carry out automatically a set of text processing operations such as hyphenation, spelling correction, grammar and style correction, summarization, and translation. So far, Neurolingo has developed the following proofing tools for M. Greek, based on the language resources mentioned above:

a) A spelling checker which flags words with spelling errors and suggests orthographically correct alternatives;

b) A hyphenator which indicates the hyphenation points of a word so as to help typesetting systems to correctly hyphenate words near the paragraph border;

c) A thesaurus which suggests synonyms and antonyms for more than 22,000 Greek words, independently of their inflection.

These three proofing tools are available for the following applications: MS Office, Open/Star/Neo Office, Quark Xpress, and Adobe Creative Suite – Acrobat, Illustrator, In Design, Flash, Photoshop and Dreamweaver – for Windows and Mac OS X.

3.2. Web-based tools

The LEXISCOPE[4] is a web-based viewer which provides grammatical and semantic information about a M. Greek word or phrase, combining Neurolingo's hyphenator, morphological lexicon and thesaurus. It also incorporates the functionality of the spelling-checker and the lemmatizer, in order to suggest orthographically correct alternatives in case of a misspelled search term, and access the information through any of its word forms.

[3] An LR parser is a parser that reads input from Left to right (as it would appear if visually displayed) and produces a Rightmost derivation.

[4] http://www.neurolingo.gr/en/online_tools/lexiscope.htm

3.3. Lexicon browser

The M. Greek thesaurus is also available as a standalone tool with the appropriate browser. As it incorporates the functionality of the lemmatizer, lexical information is accessible through any morphological form of the searched word/phrase.

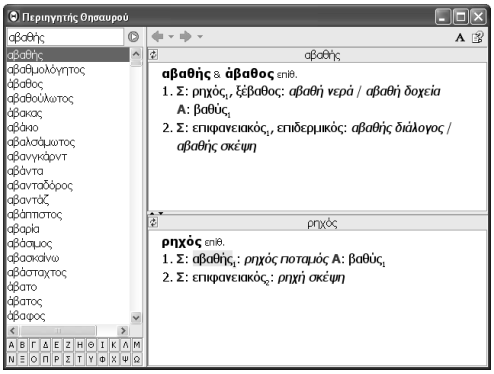


Figure 2. The Thesaurus browser

4. Avenues for future research

What normally comes next is the development of a grammar checker for M. Greek. We have already started studying and analyzing the grammatical errors in running texts and designing algorithms to handle them. Hopefully, a grammar checker prototype will be available next year. This task will be carried out through the MNEMOSYNE® system, which is an integrated document collection processing environment. The system supports different formats of input documents (*e.g.* html, pdf, text), stored in various media (*e.g.* files, web pages, databases). It incorporates the existing language resources, the KANON formalism, different types and flows of analyzers, and fuzzy matching techniques. So far, it has been used for Named Entities Recognition in the framework of R&D projects for text mining and information extraction from free text documents with very good scoring as regards the size of the input data, the speed of the processing and the output accuracy. For instance, after processing a collection of 28,000 documents in less than 4 hours, an accuracy of 90% on the recognition of ~500,000 events and ~2,600,000 Named Entities was achieved.

The language tools presented above have been used so far for the development of language resources, incorporating a number of language specific features (*e.g.* hyphenation rules, inflectional paradigms, spell-checking functionality, and syntactic patterns) to support the description of the M. Greek language. However, all of these systems are based on XML and Unicode standards and could, with minor adjustments, be used for encoding data from other languages.

References

IORDANIDOU, A., PANTAZARA, M., MANTZARI, E., ORPHANOS, G., VAGELATOS, A. and PAPAPANAGIOTOU, V. (2007). Ζητήματα αναγνώρισης πολυλεκτικών όρων στον τομέα της βιοϊατρικής [Issues of multi-word terms' recognition in the biomedical domain]. In *Proceedings of the 6th Conference on Greek Language and Terminology* (ELETO), Athens, 1-3 November 2007: 178-191.

TSALIDIS, C., ORPHANOS, G., MANTZARI, E., PANTAZARA, M., DIOLIS, C. and VAGELATOS, A. (2007). Developing a Greek biomedical corpus towards text mining. In *Proceedings of Corpus Linguistics Conference 2007*, Birmingham, UK, 27-30 July 2007. Available from www.corpus.bham.ac.uk/conference2007.

TSALIDIS, C., ORPHANOS, G., IORDANIDOU, A. and VAGELATOS, A. (2004). Proofing Tools Technology at Neurosoft S.A. Paper presented at the *Workshop on International Proofing Tools and Language Technologies*, Patras, 1-2 July 2004.

TSALIDIS, C., VAGELATOS, A. and ORPHANOS, G. (2004). An electronic dictionary as a basis or NLP tools: The Greek case. Paper presented at the *11th Conference on Natural Language Processing* (TALN'04), Fez, 19-22 April 2004.

TSALIDIS, C., ORPHANOS, G., and IORDANIDOU, A. (2002). Οι ελληνόγλωσσοι υπολογιστές έμαθαν να συλλαβίζουν (;) [Did the Greek computers learn how to hyphenate (?)]. *Studies in Greek Linguistics, Proceedings of the 23rd Annual Meeting of the Department of Linguistics*, Aristotle University of Thessaloniki, Thessaloniki, 901-911.

Cahiers du Cental

La collection « Cahiers du Cental » est une publication du Centre de traitement automatique du langage de l'Université catholique de Louvain
http://www.uclouvain.be/cental

Hors-série

Purnelle G., Fairon C. et Dister A. (éds) (2004), *Le poids des mots, Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*, 2 vols, Presses universitaires de Louvain, Louvain-la-Neuve, 1219 p.

Cahiers du Cental

1. Didier J.-J., Hambursin O., Moreau Ph. et Seron M. (éds) (2006), *« Le français m'a tuer », Actes du colloque L'orthographe française à l'épreuve du supérieur*, Bruxelles, 27 mai 2005, Presses universitaires de Louvain, Louvain-la-Neuve, v-113 p.
2. Mertens P., Fairon C., Dister A. et Watrin P. (éds) (2006), *Verbum ex machina, Actes de la 13e conférence sur le Traitement Automatique des Langues Naturelles* (2006), Leuven, 10-13 avril 2006, Presses universitaires de Louvain, Louvain-la-Neuve, 2 vols, xviii-951 p.

3.1. Fairon C., Klein J.R. et Paumier S. (2006), *Le langage SMS. Étude d'un corpus informatisé à partir de l'enquête « Faites donc de vos SMS à la science »*, Presses universitaires de Louvain, Louvain-la-Neuve, viii-123 p.

3.2. Fairon C., Klein J.R. et Paumier S. (2006), *Le corpus SMS pour la science. Base de données de 30.000 SMS et logiciel de consultation*, CD-Rom, Presses universitaires de Louvain, Louvain-la-Neuve, v-44 p.

4. Fairon C., Naets, H., Kilgarriff A. et de Schryver G.-M. (éds) (2007), *Building and Exploring Web Corpora, Proceedings of the 3rd Web as Corpus Workshop, Incorporating Cleaneval*, Presses universitaires de Louvain, Louvain-la-Neuve, viii-167 p.
5. Constant M., Dister A., Emirkanian L. et Piron S. (éds) (2008), *Description linguistique pour le traitement automatique du français*, Presses universitaires de Louvain, Louvain-la-Neuve, v-246 p.
6. Nakamura T., Laporte É., Dister A. et Fairon C. (éds) (2010), *Les tables. La grammaire du français par le menu. Mélanges en hommage à Christian Leclère*, Presses universitaires de Louvain, Louvain-la-Neuve, xvii-383 p.
7. Granger S. et Paquot M. (eds) (2010), *eLexicography in the 21st century : New challenges, new applications, Proceedings of eLex 2009*, Louvain-la-Neuve, 22-24 October 2009, Presses universitaires de Louvain, Louvain-la-Neuve, ix-462 p.