ORGANIZING COMMITTEE

De Cock Sylvie (Facultés Universitaires Saint-Louis & CECL, UCLouvain, Belgium)
Granger Sylviane (CECL, UCLouvain, Belgium)
Paquot Magali (CECL, UCLouvain, Belgium)
Rayson Paul (UCREL, Lancaster University, Great Britain)
Tutin Agnés (LIDILEM, Université Stendhal-Grenoble 3, France)

LOCAL COMMITTEE

Gilquin Gaëtanelle (CECL, UCLouvain, Belgium)
Goossens Diane (CECL, UCLouvain, Belgium)
Gouverneur Céline (CECL, UCLouvain, Belgium)
Hugon Claire (CECL, UCLouvain, Belgium)
Lefer Marie-Aude (CECL, UCLouvain, Belgium)
Litré Damien (CECL, UCLouvain, Belgium)
Meunier Fanny (CECL, UCLouvain, Belgium)
Schutz Natassia (CECL, UCLouvain, Belgium)
Thewissen Jennifer (CECL, UCLouvain, Belgium)

SCIENTIFIC COMMITTEE

Bogaards Paul (Leiden University, The Netherlands)
Bouillon Pierrette (ISSCO, University of Geneva, Switzerland)
Campoy Cubillo, Maria Carmen (Universitat Jaume I, Spain)
Cowie Anthony (University of Leeds, Great Britain)
de Schryver Gilles-Maurice (Ghent University, Belgium)
Drouin Patrick (Observatoire de Linguistique Sens-Texte, Université de Montréal, Canada)
Fairon Cédric (CENTAL, UCLouvain, Belgium)
Fellbaum Christiane (Princeton University, United States)
Fontenelle Thierry (Microsoft Natural Language Group, United States)
Glaros Nikos (Institute for Language and Speech Processing, Greece)
Grefenstette Gregory (EXALEAD, France)
Hanks Patrick (Masaryk University, Czech Republic)
Kilgarriff Adam (Lexical Computing Ltd, Great Britain)
Korhonen Anna (University of Cambridge, Great Britain)
Herbst Thomas (Universität Erlangen, Germany)
Lemnitzer Lothar (Universität Tübingen, Germany)
Moon Rosamund (University of Birmingham, Great Britain)
Ooi Vincent (National University of Singapore, Republic of Singapore)
Pecman Mojca (Université Paris Diderot - Paris 7, France)
Piao Scott (The University of Manchester, Great Britain)
Rayson Paul (UCREL, Lancaster University, Great Britain)
Ronald Jim (Hiroshima Shudo University, Japan)
Sierra Martínez Gerardo (GIL, Universidad Autónoma de México, México)
Smrz Pavel (Brno University of Technology, Czech Republic)
Sobkowiak Włodzimierz (Adam Mickiewicz University, Poland)
Tarp Sven (Centre for Lexicography, Aarhus School of Business, Denmark)
Tutin Agnes (LIDILEM, Université Stendhal-Grenoble 3, France)
Verlinde Serge (Katholieke Universiteit Leuven, Belgium)
Yihua Zhang (Guangdong University of Foreign Studies, China)
Zock Michael (CNRS - Laboratoire d’Informatique Fondamentale, France)
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
TABLE OF CONTENTS

Keynote Papers

Heid Ulrich
Aspects of Lexical Description for Electronic Dictionaries 1

L’Homme Marie-Claude
Designing Specialized Dictionaries with Natural Language Processing Techniques: A State-of-the-Art 5

Nesi Hilary
E-dictionaries and Language Learning: Uncovering Dark Practices 7

Rundell Michael
The Road to Automated Lexicography: First Banish the Drudgery... then the Drudges? 9

Vossen Piek
From WordNet, EuroWordNet to the Global Wordnet Grid 11

Papers – Posters – Software Demos

Abel Andrea
Towards a Systematic Classification Framework for Dictionaries and CALL 15

Alonso Ramos Margarita, Wanner Leo, Vázquez Veiga Nancy, Vincze Orsolya, Mosqueira Suárez Estela & Prieto González Sabela
Tagging Collocations for Learners 19

Alonso Ramos Margarita & Nishikawa Alfonso
DiCE in the Web: An Online Spanish Collocation Dictionary 23

Baines David
FieldWorks Language Explorer (FLEx) 27

Breen James
WWWJDIC - A Feature-Rich WWW-Based Japanese Dictionary 31

Breen James
Identification of Neologisms in Japanese by Corpus Analysis 35

Cartoni Bruno
Introducing the MuLexFoR : A Multilingual Lexeme Formation Rule Database 39

Corino Elisa & Onesti Cristina
Have I Got the Wrong Definition of...? How to Write Simple Technical Definitions on the Basis of Examples Taken from Newsgroup Discussions 41

Cougnon Louise-Amélie & Beaufort Richard
SSLD: A French SMS to Standard Language Dictionary 43

Damascelli Adriana Teresa
Building a Bilingual Web-Glossary of Social Services Terms as Part of a Language Learning Environment 47

de Schryver Gilles-Maurice
Artificial Intelligence Meets e-Lexicography 49
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Drouin Patrick</strong></td>
<td>51</td>
</tr>
<tr>
<td>Building a Bilingual Transdisciplinary Scientific Lexicon</td>
<td></td>
</tr>
<tr>
<td><strong>Durán-Muñoz Isabel</strong></td>
<td>55</td>
</tr>
<tr>
<td>Translators’ Needs into Account: A Survey on Specialised Lexicographical Resources</td>
<td></td>
</tr>
<tr>
<td><strong>Dziemianko Anna</strong></td>
<td>57</td>
</tr>
<tr>
<td>Paper or Electronic: The Role of Dictionary Form in Language Reception, Production and the Retention of Meaning and Collocations</td>
<td></td>
</tr>
<tr>
<td><strong>Ecker Alexandre &amp; Fichtner Ralph</strong></td>
<td>61</td>
</tr>
<tr>
<td>The Lëtzeburger Online Dictionary (LOD) Workflow Management Concept, an All-Digital Approach to Lexicography</td>
<td></td>
</tr>
<tr>
<td><strong>Eyigoz Elif</strong></td>
<td>63</td>
</tr>
<tr>
<td>A Theoretically Computational Lexicon for Turkish Verbs</td>
<td></td>
</tr>
<tr>
<td><strong>Fairon Cédric, Macé Kévin &amp; Naets Hubert</strong></td>
<td>65</td>
</tr>
<tr>
<td>Creating and Exploiting RSS-Based Web Corpora with GlossaNet</td>
<td></td>
</tr>
<tr>
<td><strong>Falk Ingrid, Gardent Claire, Jacquey Evelyne &amp; Venant Fabienne</strong></td>
<td>67</td>
</tr>
<tr>
<td>A Method for Grouping Synonyms</td>
<td></td>
</tr>
<tr>
<td><strong>Flinz Carolina</strong></td>
<td>71</td>
</tr>
<tr>
<td>DIL, an Online Bilingual Specialized Dictionary of Linguistics (German-Italian)</td>
<td></td>
</tr>
<tr>
<td><strong>Gala Nuria &amp; Rey Véronique</strong></td>
<td>75</td>
</tr>
<tr>
<td>Acquiring Semantics from Structured Corpora to Enrich an Existing Lexicon</td>
<td></td>
</tr>
<tr>
<td><strong>Gasiglia Nathalie</strong></td>
<td>79</td>
</tr>
<tr>
<td>Some Editorial Orientations for a Multi-tier Electronic Monolingual School Dictionary</td>
<td></td>
</tr>
<tr>
<td><strong>Granger Sylviane &amp; Paquot Magali</strong></td>
<td>83</td>
</tr>
<tr>
<td>Customising a General EAP Dictionary to Learner Needs</td>
<td></td>
</tr>
<tr>
<td><strong>Gurrutxaga Antton, Leturia Igor, Saralegi Xabier, San Vicente Iñaki &amp; Pociello Eli</strong></td>
<td>87</td>
</tr>
<tr>
<td>Evaluation of an Automatic Process for Specialized Web Corpora Collection and Term Extraction for Basque</td>
<td></td>
</tr>
<tr>
<td><strong>Hadouche Fadila, L’Homme Marie-Claude &amp; Lapalme Guy</strong></td>
<td>91</td>
</tr>
<tr>
<td>Automatic Annotation of Actants</td>
<td></td>
</tr>
<tr>
<td><strong>Halskov Jakob &amp; Jarvad Pia</strong></td>
<td>95</td>
</tr>
<tr>
<td>Human Versus Automated Extraction of Neologisms for Lexicography - A Discussion and a System Evaluation</td>
<td></td>
</tr>
<tr>
<td><strong>Hanks Patrick</strong></td>
<td>99</td>
</tr>
<tr>
<td>Elliptical Arguments: A Problem in Relating Meaning to Use</td>
<td></td>
</tr>
<tr>
<td><strong>Herbst Thomas &amp; Uhrig Peter</strong></td>
<td>103</td>
</tr>
<tr>
<td>Valency Patterns Online – The Erlangen Valency Pattern Bank</td>
<td></td>
</tr>
<tr>
<td><strong>Hmeljak Sangawa Kristina, Erjavec Tomaž &amp; Kawamura Yoshiko</strong></td>
<td>105</td>
</tr>
<tr>
<td>Automated Collection of Japanese Examples from a Parallel and a Monolingual Corpus</td>
<td></td>
</tr>
<tr>
<td><strong>Joffe David</strong></td>
<td>109</td>
</tr>
</tbody>
</table>
Karpova Olga & Gorbunov Mikhail
Cultural Values in Learner’s Dictionary: In Search of a Model 113

Kilgarriff Adam, Kovář Vojtěch & Rychlá Pavel
Tickbox Lexicography 115

Klosa Annette
On the Combination of Automated Information and Lexicographically Interpreted Information in Two German Online Dictionaries (www.elexiko.de and www.dwds.de) 119

Krek Simon & Gantar Polona
Slovene Lexical Database for NLP and Lexicographic Purposes 121

Krisi Tomer, Siani Assaf & Kernerman Ilan
The KXD Shell 125

Krek Simon & Gantar Polona
On the Combination of Automated Information and Lexicographically Interpreted Information in Two German Online Dictionaries (www.elexiko.de and www.dwds.de) 119

Kristoffersen Jette H. & Troelsgård Thomas
Making a Dictionary without Words 127

Kuzmina Vera & Rylova Anna
Software Demonstration. The ABBYY Lingvo Electronic Dictionary and the ABBYY Lingvo Content Dictionary Writing System as Lexicographic Tools 131

Langemets Margit, Loopmann Andres & Viiks Ülle
Dictionary Management System for Bilingual Dictionaries 135

Laureys Godelieve

Lew Robert & Tokarek Patryk
Entry Menus in Bilingual Electronic Dictionaries 145

Li Hanhong
Word Frequency Distribution for Electronic English Learner’s Dictionaries Based on BNC XML 147

Luder Marc
Building an OLIF-Based Lexical Database for Representing Constructions 151

Messiant Cédric & Poibeau Thierry
Automatic Lexical Acquisition from Corpora, some Limitations and some Tentative Solutions 155

Michiels Archibald
LEXDIS, a Tool for Measuring Lexical Proximity 159

Müller-Spitzer Carolin & Mörhs Christine
The “Online Bibliography of Electronic Lexicography” (OBELEX) 163

Niestadt Jan, Tiberius Carole & Moerdijk Fons
Searching the ANW Dictionary 167

Oliveira Claudia & Peters Pam
Ontologies in the Mediostructure of LSP eDictionaries 169

Paumier Sébastien, Nakamura Takuya & Voyatzi Stavroula
UNITEX, a Corpus Processing System with Multi-Lingual Linguistic Resources 173
Pecman Mojca, Juilliard Claudie, Kübler Natalie & Volanschi Alexandra
Processing Collocations in a Terminological Database Based on a Cross-Disciplinary Study of Scientific Texts

Piotrowski Tadeusz
Mobile Dictionaries: Situations of Use

Sass Bálint & Pajzs Júlia
FDVC - Creating a Corpus-driven Frequency Dictionary of Verb Phrase Constructions

Siepmann Dirk
The Bilexicon: A New Resource for Language Learners

Sierra Gerardo, Medina Alfonso & Lázaro Jorge
Resources for Building a Lexical Knowledge Base for an Elementary Dictionary of Sexuality

Spina Stefania
The Dici Project: Towards a Dictionary of Italian Collocations Integrated with an Online Language Learning Platform

Spohr Dennis
Towards a Multifunctional Electronic Dictionary Using a Metamodel of User Needs

Steyer Kathrin & Brunner Annelen
Wortverbindungsfelder – Fields of Multi-Word Expressions

Tascovac Toma
Beyond Lexicocentrism: The Challenge of Complex Architectures in eLexicography

Tiberius Carole & Moerdijk Fons
Fine-Tuning a Common-Sense Classification in the ANW Dictionary

Tiscornia Daniela
Building Multilingual Legal Lexicons

Tittel Sabine
Dynamic Access to a Static Dictionary: A Lexicographical "Cathedral" Lives to See the Twenty-First Century – the Dictionnaire étymologique de l'ancien français

Trap-Jensen Lars, Lorentzen Henrik & Asmussen Jørg
Access to Multiple Lexical Resources at a Stroke: Integrating Dictionary, Corpus and Wordnet Data

Tsaliidis Christos, Mantzari Elena & Pantazara Mavina
NLP Tools for Lexicographic Applications in Modern Greek

Tschichold Cornelia
From Lexical Database to Intelligent Vocabulary Trainers

Tutin Agnès
Showing Phraseology in Context: An Onomasiological Access to Lexico-Grammatical Patterns in Corpora of French Scientific Writings

Verdaguer Isabel, Laso Natalia Judith, Giménez Eva, Salazar Danica & Comelles Elisabet
SciE-Lex: An Electronic Lexical Database for the Spanish Medical Community
Verlinde Serge
From Paper-Based Electronic Dictionaries to Leximats 235

Villegas Marta, Bel Nuria, Bel Santiago, Alemany Francesca & Martínez Héctor
Lexicography in the Grid Environment 239

Vitas Duško, Nakamura Takuya & Voyatzis Stavroula
Using Aligned Corpora to Construct Large-Scaled Linguistic Resources: The Electronic Dictionary of Spanish Compound Nouns 243

Volanschi Alexandra & Kübler Natalie
Building an Electronic Combinatory Dictionary as a Writing Aid Tool for Researchers in Biology 247

Wandl-Vogt Eveline
Dialect Dictionaries at a Crossroads: Multiple Access Routes on the Example of the Dictionary of Bavarian Dialects in Austria (Wörterbuch der bairischen Mundarten in Österreich (WBÖ)) 251

Željko Miran
Integration of Multilingual Terminology Database and Multilingual Parallel Corpus 255

Zock Michael & Wandmacher Tonio
Reverse Access via an Index Based on the Notion of Association. What Do Vector-Based Approaches Have to Offer? 257
Section 1

KEYNOTE PAPERS
Aspects of Lexical Description for Electronic Dictionaries

Ulrich Heid
Universität Stuttgart - Institut für maschinelle Sprachverarbeitung
heid@ims.uni-stuttgart.de

Printed monolingual definition dictionaries are primarily structured as lists of lists (cf. Tarp, 2006): each entry is a list of items (“Angaben” in Wiegand's sense), i.e. of lexicographical data allowing the user to infer information about different linguistic properties of the lexical object treated in the entry: for example its morphological, syntactic, semantic and pragmatic properties (microstructure). These entries are themselves listed, e.g. alphabetically (macrostructure), and some of them are related by links. The links are usually only of few types: synonymy, antonymy, an unspecific “see also”-relation, etc. Electronic dictionaries which have been transformed from printed ones tend to conserve this linear structure, possibly with the inclusion of general links from each word form appearing in one of the items to that word's entry.

The claim underlying this paper is that electronic dictionaries could do better: they would serve users in different situations much better if they included more relations, and if they were typed, i.e. based on a typology of lexical objects, properties and relations. In addition to browsing use, this would support more focused search in the dictionary, as well as non-standard access to lexicographic data (i.e. access not based on a single lemma).

We will illustrate our claim by outlining partial descriptive models of the intended kind for certain types of multiword expressions and for marked vocabulary: we look at noun+verb collocations (pay attention; a question arises), at German predicative PPs in copula constructions (er ist aus dem Häuschen (“he is excited”), das ist an der Zeit (“it is time for this”)), and at geographically, domain-wise or otherwise marked vocabulary.

To treat the latter, among others relations between marked and unmarked lexical objects are necessary, and likely also directed links (e.g. from dispreferred to preferred items). For multiword expressions, two sets of properties and relations are needed: (i) those that apply to elements of the multiword, and (ii) those that concern the multiword as a whole. Deciding to use type (ii) implies that multiword expressions get the same status (of treatment units, cf. Heid & Gouws, 2006) as single word lexical objects.

A typed dictionary of the above kind will enhance access possibilities, and it allows for a simple integration of function-specific views (in the sense of Bergenholtz and Tarp's function theory): we will provide small sample fragments of German noun+verb collocations and of predicative PPs, along with function-specific views, for text understanding, for text production, and for specific production-related search. These views should be created by applying constraints on the selection of lexicographic data (filters: which types of properties and relations are relevant for a user in a given situation?) and on its presentation (sequencing, layout, etc. on screen).

With the proposed model in mind, we finally briefly analyse existing online dictionaries (an online portal, and the learner's dictionaries ELDIT (Abel & Weber, 2005 etc.), BLF (Verlinde et al, 2006) and DICE (www.dicesp.com)), as well as current representational proposals (e.g. ISO-1951 Lexical Systems (Polguère, 2006),
and LMF (Francopoulou et al., 2006)). The portals analysed (e.g. StarDict (http://stardict.sourceforge.net)) mainly reproduce the printed dictionaries, enable parallel access to all included dictionaries, and add cross-references from word forms in the article text to the respective entries; ISO-1951 is also mainly focussed on the reproduction of printed material, providing a meta-representation general enough to host quite different print dictionary formats; Lexical Systems is another format for the cohabitation of data from different sources; and LMF provides a general meta-model for dictionary data for NLP, covering a broad range of dictionary types. The electronic learner's dictionaries ELDIT and BLF are designed to provide a substantial amount of different relations between lexical objects. Nevertheless, they seem mainly to be made for browsing, but not for focused search, as their query support is still mainly lemma-based.

As far as implementations of our typed dictionary model are concerned, both relational (or object) databases and typed formalisms like OWL-DL (cf. Bechhofer et al., 2004) seem to be appropriate. Spohr (2008), for example, is working with OWL-DL. In the eLexicography conference, he presents a prototype of a multifunctional dictionary based on OWL-DL which accounts for different users' needs (Spohr, 2009). We see an interesting research potential in combining available representation systems (like OWL-DL) with a detailed enough lexicographic data description to account for a wide variety of usage scenarios.

References


leksikografisk teori med saerlig henblik på lørnerleksikografi, Habilitation, Dr. ling. merc., (Århus: Handelshøjskolen i Århus), 2 vols., 2006; English Version: 

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 (compiling corpora, finding relevant terms in these corpora, locating data that can help process terms, inserting the information collected during the previous steps in a record, updating of records, etc.), the way in which the data is handled is completely different.

In this talk, I will present a methodology for compiling an online specialized dictionary that incorporates natural language processing applications. The dictionary considered is representative of a new generation of specialized dictionaries which aim to give users access to rich linguistic information based mostly on information collected from specialized corpora. These reference works differ from most specialized dictionaries which aim at providing users with explanation on concepts similar to that given in encyclopaedias. The dictionary I will present includes terms related to computing and the Internet and provides for each of them: fine-grained semantic distinctions, argument structure, combinatorial possibilities of terms with other terms of the domain, lists of lexical relationships (e.g., synonyms, antonyms, hyperonyms, collocates), etc. The dictionary also provides syntactic and semantic annotations of contexts in which terms appear.

First, the six basic steps of the methodology will be described: 1. compilation of corpora; 2. identification of relevant terminological units; 3. collection of data from corpora; 4. analysis of the data collected; 5. compilation of term records; 6. establishment of relationships between terms records. I will proceed to show some resources and tools that can assist terminologists during some of these steps and present some of the challenges that their introduction in terminology work has raised. I will focus on: a. management of corpora in electronic form for terminology purposes; b. annotation of corpora (part-of-speech tagging and lemmatization); c. term extraction; d. automatic or semi-automatic identification of information on terms in corpora, especially for finding semantic relationships (e.g. hyperonymic relationships, collocations, or predicate-argument relationships); e. formalisms for encoding terminological data. The point of view taken when presenting computer applications will be that of users rather than that of developers.

Then, I will proceed to illustrate how other resources and computer applications can assist terminologists carrying bilingual terminology work. These applications include (in addition to those reviewed for monolingual specialized dictionary compilation): a. bilingual corpora; b. bilingual term extraction; c. comparing term extraction results between languages. Specific challenges posed by
these techniques will be discussed.

Interestingly, computer applications and the use of electronic corpora have changed the way terminologists consider specialized data and have led to the definition of new practices. This, in turn, has raised theoretical issues in terminology theory. I will look at some examples and examine their implications for terminology as a set of practices, but also as a discipline providing a theoretical framework for these practices. These changes have also created a need for terminologists with increased knowledge in natural language processing techniques, corpus-based linguistics, and lexical semantics. I will also examine some implications these changes have on training in terminology.

References (selected list)


E-dictionaries and Language Learning: Uncovering Dark Practices

Hilary Nesi
Coventry University, UK
h.nesi@coventry.ac.uk

It makes sense for teachers to find out “exactly what their students are doing with their dictionaries, what they expect from them, and how easily they are satisfied during the process of consultation” (Atkins & Varantola, 1998: 115). Such information is important for the development of dictionary skills training programmes, and ought to find its way back to publishers to inform the design of future dictionaries. In the case of e-dictionaries, however, it is difficult to uncover the facts about user habits and preferences. E-dictionary users are inclined to be secretive, especially when they sense that their teachers disapprove of the types of e-dictionary they like best. It is almost impossible to track their consultation processes when their e-dictionaries are stored on their home computer hard-drives, or when they use handheld e-dictionaries covertly in class, the screen display only visible to themselves.

Generally language learners have three purchase options when choosing an e-dictionary: they can buy a monolingual or a bilingual dictionary on disk (to store on their computer hard-drive, or on cd-rom) or they can buy a suite of bilingual and monolingual dictionaries contained within some sort of handheld device - the “pocket electronic dictionary” or PED. Disk-based e-dictionaries are the most accessible to teachers and academics. They are produced by publishing houses which provide information about the contents and provenance of their source material, and they are likely to be reviewed from a lexicographical perspective, rather than in terms of technological innovation. PEDs, on the other hand, come with next to no documentation. Manufacturers rather than publishers take responsibility for their marketing, and new models are launched and old models are withdrawn with such regularity that even if all the students in the same class bought the same brand of PED, it might turn out that they all had access to different dictionary material and different user functions. The relationship between PED content and hard-copy dictionary content is often so unclear that it is difficult for even the most energetic researcher to discover how much abridgement has taken place, and exactly which PED contains the content of what dictionary, in which edition.

Under these circumstances it is not surprising that disk-based e-dictionaries receive more attention from reviewers and the stronger approval ratings from teachers. Very little is known, however, about the extent to which some of the most highly-regarded e-dictionaries on disk are actually used. They do not seem to sell particularly well, driving publishers to include “free” cd-roms inside the covers of the corresponding hard-copy dictionaries. This way a student who buys a print dictionary ends up with an e-dictionary as well, although findings from a recent survey of 1211 Thai undergraduates (Boonmoh & Nesi, 2008) suggest that many English learners in Thailand, at least, do not even bother to explore the contents of this kind of dictionary package. Only 28% of the survey respondents claimed to own a dictionary on cd-rom, despite the fact that most of them seemed to have followed university recommendations and bought the Longman Active Study Dictionary, with cd-rom.
attached. An earlier study (Nesi, 2003) recorded the same sort of indifference to monolingual dictionaries on disk when 32 advanced learners of English were given their own personal copies of the Macmillan English Dictionary on cd-rom to use in their own time, outside class. Very few accessed it at all, but perhaps because they were conscious of language teachers’ antipathy towards other types of e-dictionary (documented, for example, by Taylor & Chan, 1994, Koren, 1997, Deng, 2005 and Stirling, 2005) most were initially reluctant to talk about the ones they really preferred to use - pocket electronic dictionaries and bilingual dictionaries on the web and on disk. Their e-dictionary use was a sort of guilty secret which they did not equate with the ‘proper’ use of print dictionaries in the classroom. Despite having received extensive training in conventional dictionary skills they failed to transfer their dictionary awareness from the page to the screen.

This paper draws attention to the kind of defects that are still prevalent in many bilingual e-dictionaries, for example the hugely popular Dr Eye and Jinshan Ci Ba, and explores the strengths and weaknesses of some of the methods that have been used to investigate e-dictionary use, such as questionnaires, keystroke logging, retrospective interviews and think aloud protocols. It identifies the need to gather and disseminate more information about the range and capabilities of commercially available e-dictionaries, so that language learners, language teachers, and the general public can make sensible choices about what kind of e-dictionary to use, according to the demands of the task and the constraints of the context.

References


Koren, S. (1997). Quality versus convenience: Comparison of modern dictionaries from the researcher’s, teacher’s and learner’s points of view. TESL-EJ, 2,3.


The Road to Automated Lexicography: 
First Banish the Drudgery... then the Drudges?

Michael Rundell
Lexicography MasterClass Ltd
michael.rundell@lexmasterclass.com

Dr. Johnson’s well-known definition of the lexicographer – a harmless drudge that busies himself in tracing the original, and detailing the signification of words – is mirrored in his dictionary’s subtitle:

In which words are deduced from their originals, and illustrated in their different significations by examples from the best writers.

Descriptive lexicography no longer insists on sourcing its data from “the best writers”, but that’s the only significant difference between then and now. Johnson has identified the three main tasks in traditional lexicography:

1. deducing the meaning of words from “their originals” (that is, from the evidence of words in use)
2. detailing their “signification”
3. illustrating them with examples

In Johnson’s time this process really did involve a lot of drudgery. The first of his tasks, for example, analysing the linguistic evidence, presupposes an earlier stage in which the data is collected – and data collection was a labour-intensive business until very recently. Think not only of the hundreds of volunteer readers whose hand-gathered, painstakingly-transcribed citations underpin the OED, but also of the heroic efforts of early corpus-builders at Brown or Birmingham, whose ambitions always outran the available technology. And corpus creation merely provides us with the raw materials for a dictionary. Each subsequent stage in the dictionary-making process involves a mixture of routine administrative tasks – the “drudgery” that Johnson complained of – and creative thinking, applied first to data analysis and then to entry-writing. What has changed since the 1960s, when Laurence Urdang pioneered the use of computers in lexicography (Hanks, 2008; Nesi, 2009) is the balance between these two elements, with machines taking on more of the routine jobs such as checking cross-references (and generally doing them better than humans).

It is convenient to think of subsequent developments in terms of two kinds of outcome:

- technologies that have enabled us to do the same things we did before, but more efficiently and more systematically.
- “game-changing” developments that have expanded the scope of what dictionaries can do and (in some respects) changed our view of what dictionaries are for

Innovations in the first of these categories have helped to improve dictionaries and make them more internally consistent, while also releasing lexicographers from a great deal of drudgery. The second category is more interesting. Large corpora and
sophisticated querying tools have spawned the new discipline of corpus linguistics, and this has led to a re-evaluation of how language works – with inevitable consequences for dictionaries. It is already a given that the object which German speakers call a *Wörterbuch* is no longer necessarily (or even primarily) a “book”. What is less obvious is that the currency of dictionaries is no longer just “words” in isolation: they now also deal with bigger language systems and syntagmatic networks.

Which brings us to the question posed in this paper’s title: given that computers have gradually taken over many lexicographic tasks which were previously performed by humans, is it plausible to foresee this process continuing to a point where lexicographers are, ultimately, replaced by machines? Back in 1998, Greg Grefenstette asked a similar question: “Will there be lexicographers in the year 3000?” (Grefenstette, 1998). He showed how a series of computational procedures – some already in place, some on the horizon, and each “bootstrapping” from the previous one in the sequence – had the potential to progressively reduce the need for human intervention in lexicographic processes. In some respects, Grefenstette’s timescale looks over-optimistic – on present form, it’s unlikely human beings of any sort (let alone lexicographers) will still be around a thousand years from now. But from a technical point of view, the future he envisaged is already close at hand.

Grefenestette belongs to the same community (computational linguists) as my long-term collaborator Adam Kilgarriff, who has made innovative technical contributions to a number of projects I have been involved in over the past 15 years or so (see e.g. Kilgarriff & Rundell, 2002). There has been plenty of trial and error, but the overall effect has been to transfer many lexicographic tasks from human to computer. This is inherently interesting (and challenging), and is also a good way of making yourself popular with budget-holders (who save money and usually get a better product). But it raises questions about the future of lexicography: will mechanization lead to lexicographers becoming deskilled? Is there still a role for informed human judgement? Have I been colluding in a process that will lead to my own redundancy? And so on.

This paper will survey a number of technologies that have been applied in the last 15 years to one or more of the key lexicographic tasks. It will conclude by speculating on what the end-point of this process might be (or indeed, whether there is an end-point at all).

References


From WordNet, EuroWordNet to the Global Wordnet Grid

Piek Vossen
Vrije Universiteit Amsterdam, The Netherlands & Irion Technologies
p.vossen@let.vu.nl

In this presentation, I will give an overview of the English Wordnet and EuroWordNet and sketch the perspective of the future of the Global Wordnet Grid. English Wordnet is the resource that is used mostly in language technology. It had and still has an enormous impact in the development of language technology. The English Wordnet is organized around the notion of a synset, which is a set of synonymous words and expressions in a language. Each synset represents a concept, and lexical semantic relations, such as hyponymy and meronymy, are expressed between synsets. Wordnet thus deviates from the traditional lexical resources that take individual word meanings as a basis.

EuroWordNet not only extended the model to other languages but also added a cross-lingual perspective to lexical semantic resources. In EuroWordNet, the synsets of all languages have been related through equivalence relations to the synsets from the English wordnet, which thus functions as an interlingua between the different languages. Through English, the vocabulary of any language can be mapped to any other language in the model. This raises fundamental issues about the cross-lingual status of the semantic information, as a system of a language or a system of knowledge of the world. Many lexical resources duplicate semantic information and knowledge that is not specific to a language. This not only leads to inconsistencies across resources but also complicates specifying the relations across the vocabularies of different languages.

This matter is taken a step further in the Global Wordnet Grid, where wordnets are related to a shared ontology that makes a common world knowledge model explicit. An ontology as an interlingua has many advantages to using a real language:

1. specific features of English, both cultural and linguistic, are not complicating the definition of the equivalence relations of languages to the index;
2. concepts that do not occur in English can easily be added to the ontology;
3. the meaning of the concepts can be defined by formal axioms in logic;
4. it will be possible to make a more fundamental distinction between knowledge of the world and knowledge of a language;
5. the ontology can be used for making semantic inferences by computer programs in a uniform way despite the language that is linked to it.

Obviously, developing such an ontology and defining the mappings from the vocabularies to the ontology will be a long and painstaking process. In recent years though, a lot of progress has been made in the area of ontology development, which makes such an enterprise more realistic. Many ontologies and semantic lexicons have been developed and represented more and more in standardized formats. Proposals are being developed in ISO working groups how to structure each and how to relate lexicons to ontologies. Distributed resources are published on the web and are intensively used both in the Semantic Web 2.0 community of social networks and the Semantic Web 3.0 community of active knowledge repositories. The time is therefore
ripe for a project such as the Global Wordnet Grid.

A first implementation of the Global Wordnet Grid is built in the current FP7 project KYOTO. An important feature of KYOTO is that the building of wordnets and the ontology is done by communities in specific domains through a Wiki environment. In this environment, the people in these communities discuss and define the meanings and concepts of the terms in their field, even across languages. As a starting point, the Wiki environment is pre-loaded with terms that are automatically derived from documents that can be uploaded. The rich term database with pointers to textual occurrences of these terms will make it easier to define the meanings in a formal way. The Wiki uses textual examples and paraphrases in interviews to validate the relations and formal definitions. The derived knowledge structures are hidden to the user but can directly be applied by other computer programs to mine important facts and data from the sources provided by the community. KYOTO is therefore not just another Wikipedia but a platform for defining and anchoring meaning across languages and across people and computers. Likewise, KYOTO allows communities to build lexicons as a form of knowledge and language acquisition from which they directly benefit to handle knowledge and facts. Since built lexicons and ontologies are also anchored to generic wordnets and a generic ontology, the distributed community effort will eventually lead to the development of the Global Wordnet Grid.

Links

English WordNet : http://wordnet.princeton.edu
Global WordNet Grid : http://www.globalwordnet.org/
KYOTO : http://www.kyoto-project.eu/
Section 2

PAPERS – POSTERS – SOFTWARE DEMOS
Towards a Systematic Classification Framework for Dictionaries and CALL

Andrea Abel
European Academy Bolzano/Bozen (EURAC)
andrea.abel@eurac.edu

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 on the one hand, and a lot about CALL has been written during the last decades on the other, there is little written specifically about dictionaries in CALL systems.

In this regard there is an obvious link to the concept of a 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.

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.

Therefore, this paper 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.

The proposed classification is along three dimensions:

The first dimension can best be captured by the notions 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. 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) or in intelligent environments (e.g. ALFALEX, ELDIT). On the other hand, the dictionary creation itself can be the aim of the system through which language learning, especially vocabulary learning, can be supported (e.g. LogoTax).

The second dimension of dictionaries used in CALL is the human- vs. machine-oriented distinction. Human-oriented dictionaries are those that are usable through a graphical user interface (GUI), and the interface design is crucial. However, the machine-oriented lexica are based on a complex internal representation the system accesses, e.g. for facilitating dictionary lookup, for error diagnosis and error feedback.
Furthermore, dictionaries can contain both elements (e.g. ELDIT). Hence, this can be interpreted as a new facet of multifunctionality of a dictionary (cf. Heid, 2006: 981 as to the notion of multifunctionality; Abel, 2003: 537 as to the flexibility and modularity of new electronic dictionaries as a basis for including them in multiple environments).

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

While CALL-applications (here: those including a dictionary component) are obviously systems especially developed for language learning, in addition, there are also systems, which are inherently intended for other purposes, but can, at the same time, be used for language learning. For instance, there are research tools which are partly experimental, partly fully operational systems (e.g. Laufer & Hill, 2005: “Words in your ear” for the investigation of dictionary lookup patterns; Chun & Plass, 1996: analysis of the effects of multimedia annotation on vocabulary acquisition). Other programs such as collaborative dictionary writing can be used for language learning too. Furthermore, applications can be interesting which are not intended either for lexicography or for CALL, but could be fruitfully used for both, as well as for other research purposes, too, e.g. tools such as the ESP games (e.g. von Ahn, 2006, Games with a purpose: the user can practice defining terms online, results can be used for experiments in order to improve the formulation of better definitions).

In conclusion, we think that the requirements for dictionaries and CALL depend on the overall aim and approach of the whole application. Therefore, in future we are interested in examining already existing systems and hope to detect more specific indications about how to describe the crucial interdependence between system, scope and dictionary.

References


Tagging Collocations for Learners

Margarita Alonso Ramos¹, Leo Wanner², Nancy Vázquez Veiga¹, Orsolya Vincze¹, Estela Mosquera Suárez¹ & Sabela Prieto González¹

¹University of A Coruña
²ICREA and Pompeu Fabra University
lxalonso@udc.es, leo.wanner@upf.edu

The importance of collocations in second language acquisition is increasingly recognized (Lewis, 2000; Granger, 1998; Howarth, 1998; Nesselhauf, 2003, 2005; Alonso Ramos, 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 (typical) collocation errors; a general corpus is needed as source of illustration and training material.

We aim at the development of an advanced NLP-based 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, (iii) tagging the corpus.

A detailed analysis of learner corpora has proved to be essential (Dagneaux et al., 1998; Granger, 1998, 2007; Tono, 2003). Such an analysis requires a predefined error 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; Miličević & Hamel, 2007; Granger, 2007; Díaz-Negrillo & García-Cumbres, 2007); occasionally, collocation errors are also discussed referring to the POS of the collocation elements (Philip, 2007). A closer look at a learner corpus, in our case, the Corpus Escrito del Español L2 (CEDEL2) from the Autonomous University of Madrid¹, reveals however that a more detailed typology is needed. 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 post [of work]’
3. recibí un llamó 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 (such as bilingual instead of bilingüe), which we ignore, the following collocation construction errors stand out:²

¹ 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 stem from learners with intermediate or advanced level of knowledge of Spanish. For more information, see http://www.uam.es/proyectosinv/woslac/cedel2.htm.
² We interpret collocations in the sense of Hausmann (1979) as idiosyncratic word co-occurrences consisting of a base and a collocate.
1) error in the base resulting from the projection of a word in L1 to L2: lograr [el] gol (correct: lograr [el] objetivo);

2) error in the collocate resulting from a literal translation of a word from L1 to L2: llenar [un] puesto (correct: 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: recibí un llamo de Brad (correct: recibí una llamada de Brad; or, better: me llamó Brad);

4) error in the number of the base and in the governed preposition: tienen prejuicio [por algo] (correct: tienen prejuicios [hacia algo]).

The errors are very different. Therefore, a sufficiently 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 (*hacer de cotilleos, lit ‘[to] make of gossip’ instead of cotillear ‘[to] gossip’), and inexistent single LUs used instead of collocations (*escapatarar instead of ir de escaparates, lit. ‘[to] go of window-shopping’). In the second case, we distinguish between errors concerning paradigmatic lexical selection and errors concerning syntagmatic lexical selection; the former concern the base, the second the collocate. Most of the lexical errors are literal translations from L1. Although a finer distinction is necessary later on, as a first approximation, the distinction between “transferences by importation” (recibir un llamo, lit ‘[to] receive a call’, instead of recibir una llamada) and “transferences by extension” (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.

In the class of register error, we group collocations that are inappropriate pragmatically. 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 be bilingual’).

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 analogous collocations. The most detailed and systematic semantically-oriented typology of collocations we know of are the Lexical Functions (LFs) (Mel’čuk, 1996).

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 some examples:

\[1') \text{deseo } <\text{1 Real}_1 \lograr> \text{ el } <\text{1 gol}> <\text{B error = Calque corr= objetivo}> \text{ de ser bilingüe}\]

\[2') <\text{1 llenar error}> <\text{C error = Calque corr=ocupar LF = Oper}_1> \text{ un } <\text{1 puesto}> \text{ de trabajo}\]
The tagging of the learner corpus is currently performed manually, supported by an interactive editor. We are also working on a tagger of collocations with respect to both the LF typology and the collocation error typology. The work on the LF-tagger draws upon the work described in Wanner et al. (2006).

In the full paper, the collocation error typology, the collocation error tag set and the tagging procedure will be discussed in detail.

References


Mel’čuk, I. (1996). Lexical functions : A tool for the description of lexical relations in


DiCE in the Web: An Online Spanish Collocation Dictionary

Margarita Alonso Ramos & Alfonso Nishikawa
University of A Coruña
lxalonso@udc.es

In the proposed demonstration, we will show the online (web-based) collocation dictionary of Spanish, DiCE (Diccionario de Colocaciones del Español), which 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, in the sense of Hausmann (1979), and others. DiCE is a dictionary of the kind of the BBI (Benson et al., 1986), LTP (Hill and Lewis, 1997) or the Oxford Collocations Dictionary (Crowther et al., 2002). However, unlike the English paper dictionaries mentioned, DiCE has been conceived as an electronic lexical database, a preliminary demonstration of which is available on the web (http://www.dicesp.com). As far as its theoretical orientation is concerned, DiCE draws upon the fine-grained typology of lexical functions (LFs) introduced in the Explanatory Combinatorial Lexicology (Mel’čuk et al., 1995) – although the user does not need to be aware of it (see below).

The objective of DiCE is to serve as both a look-up and a learner dictionary. Accordingly, in DiCE’s home page (see the snapshot immediately below), the user is offered to enter either the query section (button ‘Consultas’) or the learner section (button ‘Ejercicios’).

The query section allows the user to look up the definition of a lemma and perform a collocation search for the lemma either as base or as collocate. The base search supports searches with respect to two different collocation classifications: (i) the classification based on the POS of the collocate, and (ii) the LF-classification. That is, when using (i) the user can query for all collocations of the selected lemma...
with adjectives, with verbs, etc.; cf. the following snapshot that illustrates the lemma + Adjective case. In addition, the user can also query, for instance, for collocations expressing attributes of the actants of the lemma. When using (ii), the user can query either a specific LF or a semantic gloss of an LF.

In contrast to the conventional dictionaries, DiCE provides for each collocation a number of sentential examples that illustrate the use of this collocation in context. Cf. a snapshot with some collocations of ADMIRACIÓN ‘admiration’:

Most of the examples are extracted from the Corpus of the Real Academia Española (CREA, [http://www.rae.es](http://www.rae.es)) and selected manually.

The collocate appears in bold (absoluta: admiración absoluta, lit. ‘absolute admiration’, encendida: admiración encendida, lit. ‘emblazed admiration’, etc.). The snapshot also shows that for each collocate, the semantic gloss of the corresponding collocation is displayed – as, e.g., INTENSA ‘intense’ above. This proved to be a very useful feature especially for learners, who face the problem of a correct choice between collocations with potentially similar but clearly distinct meanings. Consider, for instance, INTENSA ‘intense’ vs. MÁS INTENSA DE LO CONVENIENTE ‘more intense than convenient’ vs. COMPARTIDA ‘shared’ vs. QUE DURA MUCHO ‘which lasts a long time’, etc. in the following snapshot:
When clicking on the FL-button adjacent to the glosses, the user can retrieve the name of the corresponding LF.

The semantic glosses of the collocations and the association of the glosses to LFs, makes DiCE a valuable resource for such Natural Language Processing applications as text generation.

DiCE also supports the inverse collocation search (the collocate search), which allows for the retrieval of all bases for a given collocate.

The exercise section of DiCE offers collocation production and collocation understanding exercises. The following snapshot displays some exercises for production:

<table>
<thead>
<tr>
<th>Cuestionario actual: Producción</th>
</tr>
</thead>
<tbody>
<tr>
<td>Puntos totales: 2</td>
</tr>
<tr>
<td>Porcentaje aciertos: 20%</td>
</tr>
</tbody>
</table>
| 1. Si Juan empieza a tener ganas de ir al cine, entonces a Juan ...

| 1 | le aparecen las ganas |
| 2 | le surgen las ganas |
| 3 | le entran ganas |
| 4 | le duran las ganas |

Correctión: ✔

2. Busca un verbo que exprese el sentido 'empezar a sentir' en la secuencia "Te ha ... cariño y sólo aspira a ser tu amigo"

Escreba la respuesta: [ ]

¿Has acertado? [ ] ✔ [ ] ✗

Respuesta correcta: [Comprobar]

3. Vincula los nombres con sus verbos correspondientes:

| 1 | espera || [ ] profesar |

All exercises are linked to the query section of DiCE; in its next release they will be also linked to a corpus-based collocation search engine in order to support
active learning.

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, etc.

Apart from the user mode, DiCE offers an editor mode and an administrator mode. The editor mode supports the lexicographers in their work on the extension and revision of the dictionary. The administrator mode is the mode for the administration of DiCE’s data base, user access and personal space management, and other administration related tasks.

During the demonstration, all features of the upcoming release of DiCE will be shown in detail.

**Implementation**: DiCE is maintained in a MySQL data base and is implemented in PHP using an Apache Server and the CakePHP environment.

**References**


FieldWorks Language Explorer (FLEx)

David Baines
SIL International
david_baines@sil.org

Introduction

FLEx is a software tool that manages linguistic data from initial collection through to preparation for publication. The program is specifically designed to assist with the analysis of any of the world's languages including those which have no defined orthography. A further aim of the program is to facilitate the publication and dissemination of the data and analysis. These design goals are the reason for FLEx's distinctiveness.

FLEx supports:

- SIL's Graphite technology, enabling FLEx to render most complex writing systems.
- Simultaneous collaboration over the Internet or a local network.
- The Dictionary Development Process.
- Stem-based or root-based dictionaries.
- Printing of a draft dictionary.
- A plug-in which provides dictionary typesetting facilities.
- Interlinear textual analysis, tagging and charting.
- Morphological analysis.
- A tool that creates a grammar sketch from the current analysis.
- Concordance tools.
- Bulk editing tools.
- A choice of user interface languages.

The data model

Behind the task-oriented interface lies a sophisticated object-oriented data model, involving some 10 classes and 88 fields for describing lexical information. There are a further 60 classes and 185 fields for morphological grammar information. The data model is mapped to a relational database which provides integrity and consistency of data. For example, a lexical relation need only be specified in one of the entries, and Language Explorer automatically displays the relation, with appropriate labels, in both entries. Furthermore, a lexical relation between two entries will be maintained even if the lexeme form of one of the entries is modified. Similarly, when an entry is deleted, any lexical relations in other entries are also removed. For advanced users the data-object model allows for direct manipulation of the data using Python, bypassing the restrictions and safeguards of the user interface.

The categorized entry tool is a good example of the task-orientation of the user interface. This tool allows for the rapid collection of words organized by semantic domain, supporting the dictionary development process and helps to produce a thesaurus or semantically classified dictionary.
The bulk editing tools enable very efficient ways of working with the entire lexicon at once, or on certain parts of it. Filters may be applied to several fields at once to select just those records to be edited in bulk. This can save hours of work for certain tasks.

**Dictionary publication**

There are numerous dictionary publication options which provide a fine level of control over the dictionary format. These include the option to have a root-based, or a stem-based dictionary. The choice between root-based and stem-based dictionary need not be made during the data entry phase. It can be delayed until preparing for publication. The dictionary preview reflects the current settings and single-entry preview is available as you are editing an entry. This gives immediate feedback, showing exactly how changes to an entry will affect the presentation in the published dictionary. The lexicon can be exported as LIFT (XML) and imported to Lexique Pro ready for publishing on the web.

DictionaryExpress is one plug-in that is available which provides typesetting facilities for dictionaries.

**Publication of linguistic papers**

Language Explorer includes a grammar sketch tool. This will gather together the grammatical information from the database and automatically create a document that describes the grammar. The grammar sketch makes a good draft for a published morphological grammar paper and it covers the following topics: Phonemes, Morpheme types, Word categories, Inflection, Derivation, Clitics, Compounding, Allomorphy, Features and Residue.

**Tutorials and support**

A set of introductory tutorial videos, which give a good overview of the software, are available. A comprehensive help file is installed with the program, which explains the function of each feature. Further technical documentation is also included covering topics such as importing data, the data model, and introductions to lexicography and morphological parsing. For any questions not covered in the help files, an active user group exists to which questions may be posted. Many of the developers are subscribed to the group and regularly answer questions. Active development of the program is continuing, and users are invited to request new features and submit error reports.

**Recommended system requirements**

- For XP Pro SP2 or 3: A pentium IV processor running at 1.5 GHz with 1.5 Gb of RAM
- For Vista SP 1: A pentium IV processor running at 2.2 GHz with 2 Gb of RAM

FLEx is available under the SIL Open Source License. At the time of writing FLEx is being used to analyse about 120 languages around the world.

**Summary**

Input:

- Rapid data entry and bulk editing, supporting the Dictionary Development Process.
- Input data using multiple orthographies (E.g. Roman, Traditional,
Phonetic).

- Input data using complex scripts.
- Import data from Shoebox, Toolbox, LinguaLinks, or LIFT (xml).

Processing:

- Filter the lexicon on a single field, or on several fields at once.
- Filter using regular expressions.
- Build a reversal index.
- Create interlinear texts and annotate them at word or morpheme level.
- Create a morphological grammar.
- Use the parser to test words, or whole texts, against the morphological grammar.
- Build concordances of occurrences of lexemes and morphemes in analysed texts.

Output:

- Configure, preview and print a dictionary.
- Produce a grammar sketch describing the morphological aspects of the grammar.
- Export to LIFT(xml) format.
- Export interlinear texts in various formats (OpenOffice HTML, XML).


References


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.

The dictionary service has been developed in an attempt to give expression to concepts of "tomorrow's dictionary" (Atkins, 2002) in providing a wide range of configurable features and options which go well beyond the common commercial dictionary services based on accesses to copies of published bilingual dictionaries (Kenkyusha Ltd; NTT Resonant Inc; Yahoo Japan Corporation).

The dictionary files used by the server are:

- a. the JMdict/EDICT Japanese-English dictionary (Breen, 2004a), which has about 140,000 entries;
- b. the ENAMDICT dictionary of named entities, which has over 700,000 entries
- c. the KANJIDIC kanji (Chinese character) dictionary (Breen, 2004b), which has detailed information on over 12,000 characters
- d. a collection of glossary files in fields such as life sciences, law, engineering, Buddhism, business, etc.

Entries in the dictionaries can be accessed either by the Japanese headwords (either the kanji form or the reading/pronunciation) or by words in the glosses. (Fig. i) 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, various dictionary indices, etc. 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.
One function of the service commonly used by translators is a text-glossing capability in which Japanese text is segmented and matched with dictionary entries. The segmentation and matching uses a combination of most of the dictionary files, and allows inflected forms of verbs and adjectives to be aligned with the dictionary forms. (Fig. ii)

Aspects of WWWJDIC's service which are of particular interest in CALL are:

a. the option of displaying a table of conjugations for any of the verbs or verbal nouns in the dictionary (approx. 17,000 entries). (Fig. iii)
b. animated stroke-order-diagrams for the 2,000 most common kanji. (Fig. iv)
c. links at the entry level to the Tanaka Corpus of 150,000 Japanese-English sentence pairs. (Fig. v) The Corpus can also be searched independently.
d. sound clips of the Japanese pronunciation of almost all EDICT entries

Other features of the service are:

a. a configurable interface enabling users to structure the display and enable or disable options to suit their needs;
b. multi-lingual operation. At present the main operating pages are available in either English and Japanese. Other languages can be added by extending the catalogue files, and a French interface is in preparation.
c. a restricted interface tailored for use with Japanese mobile telephones.
d. links from each entry to a range of online dictionaries, search engines, Japanese Wikipedia entries, the Japanese WordNet, etc.
e. an edit interface enabling users to provide suggestions, amendments, etc. about dictionary entries or to propose new entries.
f. an API enabling access from software and servers.

<table>
<thead>
<tr>
<th></th>
<th>Plain/Informal</th>
<th>Polite/Formal</th>
<th>Plain/Informal</th>
<th>Polite/Formal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-past</td>
<td>食べる</td>
<td>食べます</td>
<td>食べない</td>
<td>食べません</td>
</tr>
<tr>
<td>Post</td>
<td>食べた</td>
<td>食べました</td>
<td>食べなかった</td>
<td>食べませんでした</td>
</tr>
<tr>
<td>Te-form</td>
<td>食べて</td>
<td>食べまして</td>
<td>食べなくて</td>
<td>食べませんでした</td>
</tr>
<tr>
<td>Conditional</td>
<td>食べたら</td>
<td>食べましたら</td>
<td>食べなかったら</td>
<td>食べませんでした</td>
</tr>
<tr>
<td>Provisional</td>
<td>食べれば</td>
<td>食べますなら(ば)</td>
<td>食べなければ</td>
<td>食べませんなら(ば)</td>
</tr>
<tr>
<td>Potential(#) &amp; Passive</td>
<td>食べられる</td>
<td>食べられます</td>
<td>食べられない</td>
<td>食べられません</td>
</tr>
<tr>
<td>Causative</td>
<td>食べさせる</td>
<td>食べさせます</td>
<td>食べさせない</td>
<td>食べさせません</td>
</tr>
<tr>
<td>Causative/Passive</td>
<td>食べさせられる</td>
<td>食べさせられます</td>
<td>食べさせられない</td>
<td>食べさせられません</td>
</tr>
<tr>
<td>Volitional/Hortative</td>
<td>食べよう</td>
<td>食べよう[よう/こと]にしよう</td>
<td>食べまい(*)</td>
<td>食べまい[よう/こと]にしよう</td>
</tr>
<tr>
<td>Conjectural</td>
<td>食べるだろう</td>
<td>食べるでしょう</td>
<td>食べないだろう</td>
<td>食べないでしょう</td>
</tr>
<tr>
<td>Alternative</td>
<td>食べたり</td>
<td>食べましたり</td>
<td>食べなかったり</td>
<td>食べませんでしたり</td>
</tr>
<tr>
<td>Imperative</td>
<td>食べろ</td>
<td>食べなさい</td>
<td>食べるな</td>
<td>食べなさるな</td>
</tr>
</tbody>
</table>

Fig. iii: Example of verb conjugation table

Fig. iv: Example of animated stroke order display
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 and Japanese-Dutch files.

References


Identification of Neologisms in Japanese by Corpus Analysis

James Breen
Monash University, Australia
jimbreen@gmail.com

This paper describes recent work to extend some techniques reported earlier to identify and extract neologisms from Japanese texts (Breen, 2005; Breen, 2004; Kaji, Uno & Kitsuregawa). The purpose of the work is to extend the recorded lexicon of Japanese, both in free and commercial dictionaries.

Despite having a rich lexicon, the Japanese language has a noted tendency to adopt and create new words (Chen, 2002; Tsujimura, 1996). While the reasons for adopting new words are varied, there are 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 waribi (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 sekushuwaru 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. Segmentation software for Japanese typically use extensive lexicons to enable word
segments to be identified, and usually output unassociated strings of characters when words are encountered which are not in their lexicons. 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 & Matsumoto, 2004; Uchimoto, Sekine & Isahara, 2001; Utsuro, Shime, Tsuchiya, Matsuyoshi & Sato, 2007).

In the work described in this paper several broad techniques are used:

a. drawing on the fact that loanwords in Japanese are written in the *katakana* syllabary, thus enabling relatively straightforward extraction and comparison. Some processing is needed to separate out other classes of words, such as the scientific names of flora and fauna, which are also traditionally written using *katakana* (Nakazawa, Kawahara & Kurohashi, 2005).

b. mimicking the morphological process for forming abbreviations to construct potential abbreviations from known compound nouns. The candidate is then checked in a WWW-based corpus (Kudo & Kazawa) or via a WWW search engine to determine whether the potential abbreviation is used enough to warrant closer inspection.

c. a similar mimicking of the morphological process for forming compound verbs, also examining a WWW corpus to determine whether the potential verb is in regular use.

d. conducting post-analysis of the output of a morphological analyzer to detect when unknown words have been encountered. In such cases the analyzers usually just produce a string of *kanji* until they can resynchronize. These strings need careful analysis as Japanese is an agglutinative language which makes considerable use of single-character affixes. Having automatically identified potential new words by this method, they are checked in a WWW corpus to determine whether they are used elsewhere, and sample passages are collected to extract the meanings.

Two other aspects of Japanese neologisms also need to be determined: the pronunciation and the meaning.

In the case of loanwords written in *katakana* the pronunciation is clear from the syllabic text. It is also clear in compound verbs, where the pronunciation of the component verb roots is unchanged. For the *kanji* compounds the pronunciation is less clear, as many characters can have multiple pronunciations, and the voicing may change on some non-initial consonants (*rendaku*). As writers of newspaper articles and similar texts writers will often follow new or rare words with the pronunciation in parentheses, candidate pronunciations are generated and the texts examined for possible confirmation.

The meanings of neologisms which are loanwords or abbreviations can usually be reliably derived from the source words or compounds, however care is needed in the case of loanwords as a high proportion have nuances which differ from the original. In the case of other neologisms, an initial presumption about the meaning can usually be made based on the meanings of the *kanji* used, however it is important to examine the text passages in which the word appears to verify or determine the meaning (Uchiyama, Baldwin & Ishizaki, 2005). Often the arrival of a neologism will result in discussion about it in online forums and articles, and by searching for the
language patterns used in such discussions it is often possible to isolate definitions. For example, in discussion of the meaning of a word, the word in question is often followed by the particle pair *towa* (とは) (as for this word/passage/etc.), thus providing an identifiable text pattern to use in searches. There is scope for training machine-learning systems with such explanatory passages.

**References**

Introducing the MuLexFoR: A Multilingual Lexeme Formation Rule Database

Bruno Cartoni
LIMSI, CNRS & ISSCO/TIM-ETI, University of Geneva
cartonib@gmail.com

Many (bilingual or monolingual) dictionaries include morphological information in their list of entries, usually with the purpose of providing information about how to produce new words. In bilingual dictionaries, this kind of information is intended to help users to understand and coin new words in L2.

Representation of morphological processes in monolingual and bilingual dictionaries has often been criticised (Prcic, 1999; Dardano, Frenguelli et al., 2006; ten Hacken, Abel et al., 2006; Cartoni, 2008). All of these studies point out the inadequacy of using only affix representation, even though it is almost always the only possibility for including morphological elements into the dictionary list of entries. Main problematic issues come from the fact that many affixes participate in more than one word-formation process, and one affix can have more than one meaning.

In the lexematic approach to morphology (see Fradin (2003) for the most recent studies in this framework), affixes are only the formal component of Lexeme-Formation Rules (hereafter LFRs), which contain other constructional operations (change of category, semantic function) and which, most importantly, are “semantically driven”.

In this paper, we present a project that aims to build an electronic database that gathers Lexeme-Formation processes in a multilingual perspective. The MuLexFoR database (Multilingual Lexeme-Formation Rule database) will provide a new solution for representing word-formation process in lexical resources.

On the lexicographic side, adopting the lexematic approach is very useful for building multilingual LFRs that match equivalent constructional processes from different languages, according to the meaning they coin. For example, in French and Italian there is one single LFR of “reiterativity” which builds verbs on verbal bases (LFR_reiter(v→v) and uses only one prefix (ri in Italian, re in French). The semantics of the rule is used as a “pivot” in the translation process, and consequently ri and re can be theoretically seen as the surface forms of one single LFR. Another advantage of this approach appears in cases where one single rule can have more than one affix. For example, the LFR for “unspecified plurality” contains three prefixes in both languages (IT: multi, pluri, poli; FR: multi, pluri, poly), although monolingual constraints can probably be applied to select the appropriate affix.

On the implementation side, the use of a multi-access and dynamic database

---

1 Some interesting attempts to rationalise morphological information and to present it in a user-friendly way have already been implemented, but only in a monolingual perspective: E. Bernal’s DSVC, database for Catalan affixes (Bernal & DeCesaris, 2008) is, to our knowledge, the first attempt to implement a rule-based approach. This project largely inspired the multilingual database presented in this article.

2 The MuLexFoR database is implemented in PHP and will be soon available on the web.
allows the user to access morphological information through different modes and different languages. First of all, the user can browse the LFRs by semantic denomination (by choosing for example the “reiterativity rule”) thus showing the whole multilingual LFR, with the affixes and constraints present for that type in the different languages of the database. Obviously, this option requires a high level of competence in morphology, therefore, two other access modes are provided.

First, the user can choose the affix s/he wants to look for in the affix index and thus have access to all the rules in which it is involved, together with a complete description of each rule and the equivalent affixes in a target language, with their usage constraints. For example, if an Italian user wants to know how to express the prefixation in multi in French, s/he can browse the affix index and select the Italian prefix multi. MuLexFoR then provides all the rule(s) that involve(s) this Italian prefix, together with a complete description of each rule, where the user will find the French equivalent affixes (multi, pluri and poly), the constraints to use them, and some examples.

The other access mode is through the lexical index, which gives access to the rule that coined the chosen word. In the future, this lexical index can simply constitute a “bridge” between the constructed lexemes of a bilingual dictionary and the MuLexFoR database.

For the moment, the database contains more than 130 LFRs that imply the 54 most productive prefixes in French and Italian. The implementation of other languages and other word-formation elements is currently under work. Some evaluations of the use of this database are also planned. First results show some interesting issues about the coverage of the database and the information that are expected by the users.

References

Have I Got the Wrong Definition of…? How to Write Simple Technical Definitions on the Basis of Examples Taken from Newsgroup Discussions

Elisa Corino & Cristina Onesti
Università di Torino
elisa.corino@unito.it – cristina.onesti@unito.it

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, developed at the University of Turin (Corino, 2007). 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.

If we focus on the following examples, we can easily notice how people need the support of “experts” to understand the subsenses of technical terms.

(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 !!

(2) A: scusate la mia ignoranza , ma cosè il Cruise Control ?

B: Per definizione dall’inglese Cruise , è la crociera , cruiser è l’ incrociatore or 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). [NUNC-IT Motori]

It is common knowledge that technical terms often come from ordinary lexicon which is turned into more specific subsenses in specialised fields (e.g. depression in economics or meteorology). Newsgroup messages provide a sort of “natural definition” proposed by experts in such fields, who try to write in as clear language as possible, though still being very specific.

1 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.
Following Ding (2008), we intend to thoroughly investigate how definitions could be improved, pointing out errors and other problems detected in the dictionary and reported directly by users in their posts. We analyse some contexts in which people ask for explanations and we sketch a pattern of how "experts" answer questions and put their clarifications in plain words. The reply mechanism, typical of computer-mediated communication (CMC), enables us to verify the success of the explanation given by the “experts” through an immediate feedback.

We also aim to outline a sample definition for technical terms by taking into account some of the key problems of lay users facing LSP, and the strategies used by “experts” to make a term clear, both as for structure and words used.

The findings shed light on some problems of technical definitions in monolingual dictionaries and provide useful information to improve them. In addition, we suggest that the analysis of real use of words and combinations of words can suggest possible issues to build a new kind of technical dictionary based on the doubts of users on the one hand, and on more comprehensible explanations on the other.

References


SSLD: A French SMS to Standard Language Dictionary

Louise-Amélie Cougnon & Richard Beaufort
Université catholique de Louvain, Cental
louise-amelie.cougnon@uclouvain.be, richard.beaufort@uclouvain.be

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 SMS, also called text messages, on the same level as chat language, is one of these new written forms of communication. When dealing with SMS\(^1\), one has to cope with various issues: new linguistic phenomena, language processing difficulties and lexical resources limits. Linguistic phenomena in the 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\(^2\), but unfortunately, studies on transcription from SMS to standard language are few and results are still too basic (mainly because they are based on limited size corpora\(^3\)).

Since the beginning of the sms4science project\(^4\), we have been questioning the usefulness of SMS corpora and SMS to standard language transcription. At this stage, we had already worked on SMS transcription, especially on the reverse dictionary: standard to SMS language\(^5\). Then, in early 2008, a new project was set up within the framework of our research centre: Vocalise, an SMS-to-Speech synthesis project\(^6\). In order to improve SMS speech synthesis, the new project developed an SMS word alignment based on a corpus of 30,000 text messages\(^7\) and their manual transcription\(^8\). It was, for us, the opportunity to address the question of SMS to standard language transcription again. We decided to use the Vocalise aligned corpus to draw up an SMS to Standard Language Dictionary (SSLD). In order to reach this target, we built a list of entries based on all the words of the aligned corpus; each entry (in SMS language) is followed by the lemma of its standard transcription and by its codes (grammatical, semantic and inflectional\(^9\)). For ambiguous terms, a new entry was created for each possible grammatical interpretation. Then, we manually filtered out unwanted entries (deleting inappropriate interpretations, unknown words and symbols, disambiguating,

\(^1\) Short Message Service. The acronym refers to the service as much as to the messages exchanged during the service themselves.

\(^2\) They are of variable qualities but in general of bad quality.

\(^3\) Except for Guimier de Neef and Fessard’s project (Guimier de Neef & Fessard, 2007), which used a corpus of 10 000 SMS.

\(^4\) A project that aims at collecting an international SMS corpus: [www.sms4science.org](http://www.sms4science.org)


\(^6\) See [http://cental.fltr.ucl.ac.be/projects/vocalise/index_EN.html](http://cental.fltr.ucl.ac.be/projects/vocalise/index_EN.html) The first step of the project is an automatic transcription from written SMS language to written standard language; the standard corpus is then vocalized.

\(^7\) Even though it is indeed a substantial corpus, it is still restricted; hopefully, we will improve the quality (representativeness) and quantity of this corpus thanks to new corpora collection that are planned for 2009 and 2010.

\(^8\) The project “Faites don de vos SMS à la science” collected 30,000 French text messages in 2004.

\(^9\) Our system of codes is largely inspired by the DELAF dictionary syntax.
completing, etc.) and analyzed this list so as to obtain a smarter SSLD.

The purpose of our paper is, at first, to detail the different steps of our methodology (corpus collection and manual transcription, Vocalise alignment, rough SMS lexical resource, smart SSLD); secondly, we will scrutinize our results, the difficulties that appeared such as, for example, double lexical entries and ambiguities.

(1) écoute,écoute.N+Abst+z1:fs
(2) écoute,écoute.N+Conc+z1:fs
(3) écoute,écouter.V+se+p+i+E+z1:P1s:P3s:S1s:S3s:Y2s
(4) écoute,écouter.V+t+z1:P1s:P3s:S1s:S3s:Y2s

It seems legitimate to wonder whether all entries are relevant and should be kept as such, or if they should be filtered out because there is little chance for them to appear (for example, écoute.N+Abst, that stands for « guetteur, poste de guet »\(^1\)) in an SMS context.

We also deleted errors, mostly adjunctions of letters or agglutinations, due to wrong segmentations during the word automatic alignment process (énormalmt). Some of these were actually identified later as SMS users’ mistakes (devd); of course, these last forms were not considered as entries for the dictionary.

(5) énormalmt,normalement.ADV+PADV+z1
(6) devd,devoir.V+i+U+z1:I1s:I2s

Another major difficulty was the presence of unrecognized entries, such as the following:

(7) él,elle.PRO+PpvIL+z1:3fs
(8) éle,UNKNOWN
(9) éle,elle.PRO+PpvIL+z1:3fs
(10) selmen,seulement.ADV+PADV+z1
(11) slm,UNKNOWN
(12) slmt,seulement.ADV+PADV+z1

For each of these cases, we searched for the origin of the problem, often contextual, sometimes due to the initial manual SMS corpus transcription or to our methodology itself.

The fourth important task that had to be implemented was to complete the corpus: because the dictionary is based on a limited corpus, only a limited proportion of language vocabulary was extracted, and, even for a given entry, all its possible standard forms were not proposed (due to the limitations of the applied dictionary). We tried to find solutions to each of these problems.

Finally, our paper will give a panorama of SSLD applications, such as enhancing speech synthesis systems.

References


\(^1\) English translation: « lookout, watcher ». 

44
Oxford: Oxford University Press.
Building a Bilingual Web-Glossary of Social Services Terms as Part of a Language Learning Environment

Adriana Teresa Damascelli
University of Torino, Italy
adrianateresa.damascelli@unito.it

The ongoing process of renovation which has been taking place since the Bologna Agreement in 2002 is still bringing about changes in the Italian university curricula system. On the one hand, there is the need to conform to the European university range of courses offered; on the other hand, there is the need to provide the job market with new professionally qualified specialists and as a result of a diversified reality the need for collecting resources useful to establish well-programmed training courses. In fact, the already available resources do not always meet learners and teachers’ expectations, and often teachers are led to use pre-packaged materials, which do not actually and effectively respond to the contents which should be learnt and taught, and subsequently adapt them to their own needs.

Since 2005 the Language Centre for Humanities of Turin University has been in charge of the English language programme for the degree course in “Social Services” organised by the Faculties of Educational Sciences, Psychology and Medicine where future social workers are expected to achieve the B1 level (according to the Common European Framework), which corresponds in more general terms to an independent use of language.

The lack of “ad-hoc” teaching and learning materials has led to the publication of a manual of English for the Social Services (2002) covering a wide range of topics which are inherent to the field of social work (e.g. child and family support, mental illness, substance abuse and poverty and social exclusion) and providing learners with many activities aiming to reinforce grammar as well as reading comprehension skills. Moreover, since social work is a practice which benefits from several domains such as psychology, legal, educational and social sciences, many terms have been imported and others created on purpose. Developing and improving vocabulary skills is important as well. In particular, learners should be trained to identify and isolate the terms which are used to connote actions, techniques, methods, and activities related to the field. Subsequently, they should be helped observe terms in contexts and understand their meanings. Translating terms, should be, then, another step. The aim would be highlighting the shared features and differences in the British and Italian system of the social services by providing a cultural perspective.

In this paper the reader will be presented with a project which is being carried out in order to provide undergraduate learners with a reference tool enabling them to consult English documents in the field of social services and a new resource for improving their lexical skills. The first step to achieve this goal is the collection of a corpus of one million words in English. Samples of texts taken from specialised journals, magazines, newspaper articles, essays and books, as well as the Internet in the domain of social services are included. The corpus will constitute a source for term extraction and a repository of texts to be included as extra materials for the consultation of the terms in context. The second step, which is at the core of the project, is the building of an electronic bilingual glossary of terms which will be
available on CD-ROM and on the Internet, the online version being accessible through a password. In both versions, an introductory section will provide the reader with an alphabetical list of the terms included in the glossary or a starting concept map which gathers the terms according to the topics; both can be considered as a starting path to consult and get information about each term. The terms constituting the glossary have been retrieved from the collected corpus and are key terms related to the readings included in the manual currently used by learners. Each entry of the glossary will include translation, meaning and examples of usage. However, links to further reading material resulting from the adaptation of newspapers and magazine articles are included as well in order to provide learners with extra material for the identification and isolation of other terms which can be subsequently investigated in the available corpus. For the translation of each term learners can consult any electronic or paper resource. The aim is to better understand the way learners approach to term translation. Subsequently, they will be provided with a comparable Italian corpus of one million words in order to observe differences of usage in the two languages.

A further step will be the design of reading comprehension activities and exercises (e.g. filling in the gaps, matching vocabulary) to help learners practice the terms acquired. In fact, notwithstanding the availability of glossaries on the Internet, learners need to develop language usage skills, by showing their ability to produce texts and use terms appropriately.

References

Artificial Intelligence Meets e-Lexicography

Gilles-Maurice de Schryver
Ghent University, BE & University of the Western Cape, SA
gillesmaurice.deschryver@UGent.be

The future of lexicography is digital, so much is certain. Yet what that digital future will look like is far less certain. In the current paper, the Web takes centre stage, and the novel type of lexicography that is proposed revolves entirely around 'you'. What is needed is a dictionary that is truly adaptive — meaning that it will physically take on different forms in different situations, and one that would do so as intelligently as possible — meaning that it would have the ability to study and understand its user, and based on that to learn how to best present itself to that user. With this, the field has moved to a very different paradigm indeed, to that where artificial intelligence meets e-lexicography.

This paradigm shift is illustrated by means of the description of a project to compile an online Swahili to English dictionary. Swahili is both the most widely spoken African language, and the one sub-Saharan language most commonly taught throughout the world. As a theoretical framework for the development of this new type of electronic dictionary, the fuzzy answer set programming framework (Van Nieuwenborgh et al., 2007) is advanced. This framework combines the advantages of rule-based languages for knowledge representation with those of fuzzy logics, resulting in a non-monotonic, declarative formalism to model relationships between continuous variables. Thus, rather than using crisp threshold values to define terms, gradual definitions can be used. Note that the gradual nature of the conditions of these rules is consistent with the observation that also their conclusion, viz. the familiarity of a user with the language, is a matter of degree. The non-monotonic nature of (fuzzy) answer set programming also makes it an ideal framework for modelling exceptions to rules, which is one of the most straightforward and interesting ways of refining the initial rules. Also observe that fuzzy answer set programming allows for a seamless integration of qualitative and quantitative user preferences, whereas existing models tend to be suitable only for qualitative (e.g. using traditional answer set programming), or only for quantitative (e.g. using weighted constraint satisfaction) models.

Two earlier prototype online dictionaries (De Schryver & Joffe, 2003; Hillewaert & De Schryver, 2004) showed that with a well-designed tracking function, any number of individual user's look-up strategies may be monitored and logged across time, which is especially relevant for studying language acquisition aspects such as vocabulary retention, and for drawing up the user profiles needed for an adaptive and intelligent electronic dictionary (cf. De Schryver & Joffe, 2004; De Schryver et al., 2006 respectively).

When it comes to the lexicographic work proper — and building on top of TshwaneLex, the world's only truly off-the-shelf lexicography software currently available — both the theoretical framework and the logs will be kept at hand during the compilation of the dictionary. On a first axis, three levels can be recognized with regard to the contents of the database proper: (1) the DTD (document type definition), which drives the structure of each and every article, (2) the lexicographic information
which populates the database, and (3) the layout / formatting of that information. These three levels are (and have to be) strictly separated. On a second axis, and this with a 'smart' dictionary in mind, level (2) will (and has to) be divided into two further layers: (2a) attribute lists (for all metalexicographic data, i.e. all recurrent information which is best selected from closed lists), and (2b) the lexicographic data proper (i.e. the dictionary information itself, for each and every lemma compiled and derived from corpus data). On a third and final axis, provision needs to be made for the possibility that each of the previous levels is further subdivided into $N$ number of layers (with $N$ derived in bootstrap-fashion as the project proceeds). Expressed in simple terms, this procedure means that, say, instead of having one definition in the database for a particular sense, one has $N$ definitions, graded according to various parameters, chief amongst them (perceived) difficulty. Likewise, and as another simplified example, multiple types of example sentences are available, including, at one extreme, different ranges of pointers directly (in)to the attached corpus lines. During actual dictionary consultation, any particular user is of course never presented with all these layers of data, rather, the artificial intelligence modules present those and only those which the user most likely wants / needs to see at that point in time.

This project, therefore, revolves around a revolutionary type of electronic dictionary in which the potential is explored, and exemplified for a Swahili to English dictionary, to link an automatically derived dynamic user profile to the proffered multimedia lexicographic output. Such adaptive and intelligent dictionaries are, by design, excellent tools to study genuine dictionary use, which in turn leads to exciting answers to age-old as well as new lexicographic questions. The three main objectives are: (1) to undertake fundamental research into the theoretical underpinnings of adaptive and intelligent lexicography, (2) to analyze and obtain new insights into true dictionary use (i.e. unobtrusive dictionary use in real dictionaries without the manipulation of any research variables), and (3) to compile a ground-breaking reference work for Swahili. Beyond these initial goals, this project has the potential to trigger a positive theoretical and technological tsunami for many years to come, in areas as varied as mobile (phone and other) applications, reference works for any of the world's languages, as well as CALL lingware at large.

References


Building a Bilingual Transdisciplinary Scientific Lexicon

Patrick Drouin
OLST, Université de Montréal, Québec, Canada
patrick.drouin@umontreal.ca

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 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 gather 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.

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. The 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 (about 200,000 words). Although the data is not exactly the same in both languages, we were able to come up with a balanced bilingual corpus. All the theses included were published between 1997 and 2007. The preprocessing of documents was performed using freely available tools. The first step in preparing the document is handled by TreeTagger (Schmid, 2004), 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. Using such a simplification will allow us to compute frequencies on lemmas instead of inflected forms.

Corpus specificity is evaluated using a measure proposed by Lafon (1980) and called specificity test (calcul des spécificités). This measure allows us to compare the frequency of a word in a corpus (here our theses corpora or TC) to the frequency of the same word in another corpus (our reference corpora or RC). 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 TC based on a normal distribution based on the observations made in the RC. The second ones have a frequency which is lower than expected while the frequency of the last ones is in normal range. For the current research, we are solely interested in positive specificities. The French reference corpus is built from 30 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 similar to the one we used in French, we divided the BNC into genres
using David Lee’s classification (2001). Lee has divided the BNC into 46 genres and
8 super genres. For our experiment, we decided to consider 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 genres.

There are at least two ways to look at distribution when dealing with a corpus
like the one 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. The other method would consist
of making sure that the 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 simply compare the raw frequency of words across the corpus without taking
into account the size of the sub-corpora. For our experiment, which we want to be
completely independent of subject areas and for the sake of processing simplicity, we
decided to go forward with the first method.

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 TC.
In order to be included in our list, a word both needs to appear in more that 100% of
the specialized sub-corpora and to have a high-specificity level.

All results obtained are validated manually. The table below summarises, for
both languages, the number of words comprised in the raw list produced using the
method described previously.

<table>
<thead>
<tr>
<th>Part of speech</th>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nouns</td>
<td>566</td>
<td>634</td>
</tr>
<tr>
<td>Adjectives</td>
<td>397</td>
<td>355</td>
</tr>
<tr>
<td>Verbs</td>
<td>185</td>
<td>752</td>
</tr>
<tr>
<td>Adverbs</td>
<td>178</td>
<td>142</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1326</strong></td>
<td><strong>1888</strong></td>
</tr>
</tbody>
</table>

We are now looking carefully at the results in order to distinguish manually
word senses in each language and to add definitions to the lists. All data is contained
in XML structures and will be published to the Web in the near future. Once the
language specific analysis is completed, we will be able to link the two lists together
in order to present a bilingual TLS.

References

Wellington.


Gledhill, C. (2000). The discourse function of collocation in research article

français élémentaire: étude sur l’établissement d’un vocabulaire et d’une grammaire

1, 128-165.


Lexicography is facing new challenges in the 21st century and, therefore, it is being developed toward new and more appropriate applications to satisfy their users’ needs and to adapt themselves to new technologies. But how satisfied are users regarding lexicographical resources? Which level of satisfaction is managed by this kind of resources? 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.

In this paper, we will present a survey carried out to improve the development of lexicography regarding professional translators’ needs with regard to specialised lexicographical resources, which are regarded as indispensable for translators. 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 parallel corpora or from existing translations.

To the best of our knowledge, translators’ needs for this kind of resources have never been seriously taken into consideration. In fact, any 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 (East, 2008; Pujol et al., 2006; Sánchez Ramos, 2005; Bogaards, 2005; Corpas Pastor et al., 2001) but none of them have focused on professional translators.

This project seeks to fill this existing gap by identifying the real needs of translators with regards to terminology. More specifically, this paper will 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 know their expectations and needs prior to the development of such a resource. It is hoped that the identification of the user’s needs with regard to terminology could lead to useful resource development projects in the future.

This survey was designed in line with established recent survey practices (Dillman, 2007; Groves et al., 2008). It was made available to translators via specialised mailing lists (Corpora List, The Linguist, Termilat, Traducción), and through several organisations for translators and interpreters (ACT, AIETI, ASETRAD, ITI, ASTTI). It was also sent out to a number of translation companies as well as individual translators. Surveys based on feedback from specialized communities have played an increasingly important role in identifying the needs, perceptions or aspirations of representatives in a certain community. Within the community of professional translators, recent surveys of term extraction tools (Zielinski & Ramírez Safar, 2005) and of translation memories (Lagoudaki, 2006) have been well received by the scientific community and now serve as reference
points for many members of this community.

In the first part of the survey users 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 of the survey focused more specifically on various aspects of terminological resources. First, users were asked to describe what terminological resources they use (encyclopedias, dictionaries, thesauri, parallel corpora/texts etc.) and the format of these resources. Next, they were asked to elaborate on the use of resources depending on the different scenarios (reading L1, reading L2, translating into L1, translating into L2); on the way they identify terms (e.g. manually or with the help of a term extractor tool) and on the texts that use to mine automatically for terms etc. Finally, users were asked about their own assessment of the resources that they use. In doing so they were asked to consider any problems or inconveniences they may have experienced, as well as taking into account issues of presentation and information they thought should be present in an “ideal” terminological resource.

The final version of this paper and this presentation will report on (and illustrate the) various statistics associated with the results of this survey and identify the needs and report on the expectations and desiderata of translators, thereby illustrating what their “ideal” terminological resource would be.

References


Paper or Electronic: The Role of Dictionary Form in Language Reception, Production and the Retention of Meaning and Collocations

Anna Dziemianko
School of English
Adam Mickiewicz University, Poznan, Poland
danna@ifa.amu.edu.pl

The aim of the present study is to compare the usefulness of a monolingual English learners’ dictionary in electronic and paper form in receptive and productive tasks, and assess the role of dictionary form in the retention of meaning and collocations.

The study attempts to answer the following research questions:

1. Is the electronic dictionary more useful in L2 reception and production than the paper dictionary?
2. Which dictionary, paper or electronic, is a better learning tool? In other words, is the retention of meaning and collocations dependent on the form of the consulted dictionary?

The investigation centers on paper and electronic versions of a recent monolingual English learners’ dictionary, COBUILD6 (2008). It is hoped that the study can contribute to the research in the field in two ways. First, it goes beyond receptive tasks, which tend to prevail in the literature on the subject. Second, it aspires to compare long-term educational advantages of paper and electronic dictionary consultation. It is believed that this aspect of the investigation can give an insight into pedagogical implications of the choice of dictionary form. Overall, the study responds to the call for research on the role of paper and electronic monolingual English learners’ dictionaries in language learning (Nesi, 1999; 2000b). Hopefully, it may also indicate whether paper or electronic dictionaries are well worth integrating into language learning and teaching.

The experimental method

To answer the research questions, an experiment was conducted, in which 64 upper-intermediate and advanced students of English at Poznań University took part. The experiment involved a pre-test with a questionnaire, a test proper and an unexpected delayed recall test.

The test proper consisted of two tasks: one involving receptive (passive) recall and the other productive (active) recall (Laufer & Goldstein, 2004: 406; Nation, 2001: 358-360). In the former, the meaning of the following nine phrases had to be explained: backgammon, booby prize, clampdown, collateral damage, down under, dream ticket, flapjack, onus and outcrop. In the latter, nine English sentences were to be completed with appropriate prepositions removed from the following collocations: under sedation, in cahoots (with), at gunpoint, wreak havoc on, in the pipeline, in the offing, on the blink, up the creek, on the trot. The tasks were given to the subjects on paper. Each subject was assigned to work with either paper (PD) or electronic (ED) version of COBUILD6. The latter is available on-line, and can be accessed by a pin code. Importantly, the dictionary information necessary to complete the tasks is the
same in the two versions of COBUILD6, which made it possible to investigate the role of dictionary form. Before the experiment, the subjects participated in an orientation session, during which they acquainted themselves with the access structures in both dictionary formats.

In the pre-test, administered immediately before the test proper, the subjects were asked to perform the same tasks without access to dictionaries. The few cases where they proved to be familiar with the target items were later excluded from analysis.

The pre-test was accompanied by a short questionnaire, which, among other things, was to reveal whether the subjects’ preference for paper or electronic pedagogical dictionaries of English could be a factor influencing results.

Immediately after the sheets with the pre-test and the questionnaire had been collected, the test proper was administered and the subjects were asked to perform the receptive and the productive tasks using the dictionaries to which they had been assigned. They were also requested to consult the dictionaries only to complete the tasks and refrain from any lateral browsing, in case it might affect vocabulary learning. The subjects’ behavior during the experiment was carefully monitored so that no such browsing took place.

Two weeks later, an unexpected vocabulary retention test was administered. The test corresponded to the test proper but the order of the target items was changed so as to reduce the risk that the subjects would rely on their recall of the sequence of correct responses.

Results

The results of the questionnaire reveal that in each experimental group, comparable proportions of students were used to consulting electronic and paper dictionaries (PD group: $Z_{obs}=-0.447$; ED group: $Z_{obs}=-0.577$; $|Z_{crit}|=1.960$, $p<0.05$). Thus, the subjects’ preference for either dictionary form could not have affected their performance.

The results obtained in the test proper and in the delayed recall tests are given in the table.

<table>
<thead>
<tr>
<th>Task</th>
<th>Dict.</th>
<th>P1%</th>
<th>Dict.</th>
<th>P2%</th>
<th>Z Test</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>proper</td>
<td>receptive</td>
<td>92,0</td>
<td>Paper</td>
<td>98,7</td>
<td>-3,805</td>
<td>0,0001*</td>
</tr>
<tr>
<td></td>
<td>productive</td>
<td>92,2</td>
<td>Electr</td>
<td>98,5</td>
<td>-3,446</td>
<td>0,0001*</td>
</tr>
<tr>
<td>Retention</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>test</td>
<td>receptive</td>
<td>62,4</td>
<td></td>
<td>76,7</td>
<td>-3,430</td>
<td>0,0003*</td>
</tr>
<tr>
<td></td>
<td>productive</td>
<td>45,7</td>
<td></td>
<td>63,8</td>
<td>-3,810</td>
<td>0,0002*</td>
</tr>
</tbody>
</table>

In the test proper, the subjects consulting the electronic dictionary performed much better than those using the paper dictionary. The difference is highly significant for both receptive and productive tasks. The consultation of the electronic dictionary was also more beneficial to the retention of both the meaning of the target items and the target collocations. The difference between paper and electronic dictionary conditions is statistically highly significant for each task in the retention test as well.

Discussion

The study provides evidence for integrating the pedagogical dictionary of English in electronic form into language learning and teaching. The results suggest that the electronic dictionary is more useful than the paper dictionary in receptive and
productive tasks. It also proves to be a better learning tool, since it enhances both passive and active recall. Possibly, the visual impact it creates and the prominent position of a headword on the computer screen attract more users’ attention than a printed page.

References


The Lëtzebuerger Online Dictionary (LOD) Workflow Management Concept, an All-Digital Approach to Lexicography

Alexandre Ecker & Ralph Fichtner
Ministry of Culture (Luxembourg)
alexandre.ecker@lod.lu, ralph.fichtner@lod.lu

The Lëtzebuerger Online Dictionnaire is a synchronic multilingual (Luxembourgish, German, French, English, Portuguese) dictionary that corresponds to the specific linguistic situation of the Grand-Duchy of Luxembourg. Its language law (1984) recognizes three official languages: the indigenous language Luxembourgish, a predominantly oral variety increasingly used in written domains, German, the language of alphabetisation, and French, the traditional language of legislation and administration. Numerous immigrants, mainly from Portugal, have contributed to enriching the linguistic diversity of a country whose foreign residents constitute more than one third of the population.

The LOD, the only dictionary of its kind offering a Luxembourgish word list, can be accessed in its current state of progress (letters A to I) on www.lod.lu. Since first going online in 2007, the LOD portal has enjoyed considerable success and is now firmly established as the linguistic reference on the Luxembourgish language.

Contrary to other multilingual dictionaries, which are often primarily focused on an encoding task, the LOD project favours a semasiological approach, relying on the linguistic abilities of its target public, i.e. the trilingualism of the Luxembourgish-speaking public and the mother tongue of allophone users. Semantic restrictors such as hypernyms (a.), (partial) synonyms (b.) or preferred collocations (c.), which clarify semantic ambiguities, have therefore been applied to translation elements rather than to the items addressed.

(1) Bréck 1. bridge [structure] 2. bridge [gymnastic exercise] 3. …

(2) behandelen 1. to treat [to deal with] 2. to treat [to look after]

(3) adoptéieren 1. to adopt [a child] 2. … 3. to pass [a law]

The index may be used productively by the Luxembourgish-speaking public to translate Luxembourgish into one of the four other target languages, while the indexes of the translations also enable allophone users to access Luxembourgish items via their (German, French, English, Portuguese) equivalent.

For all aspects of the project, be it in the lexicographic drafting, work flow management or the circulation of content, the LOD team relies exclusively on digital technology, with XML technology constituting the foundation of the work. Articles are entirely drafted within the structured environment of an XML schema, written to answer the specific structural demands of the project.

The LOD management tool DOVE was developed internally to ensure the free flow of lexicographic work in general and the LOD project in particular. It is completely decentralized and accessible from any Internet terminal, thus guaranteeing a reliable and transparent method for the editing of files and their migration between the different stages necessary before online publication.
Because of the multiple micro-structural parts of the dictionary and notably, its numerous translation elements, the finalisation of an article necessitates the intervention of several contributors who, depending on their respective skills, share in the different stages of drafting, proof-reading and correction.

Members of the team do not necessarily work on the LOD premises. In fact, the only two members who work there on a full-time basis are the co-ordinators of the project, while others may work from and live abroad. Thus, the system in place allows truly co-operative work by profiting from the best of the two worlds of XML and database.

Once the base version of an article (Luxembourgish-German-French) has been approved by a committee of proof-readers, it can be edited simultaneously by different members. The resulting modifications will then be merged to form a final, complete and corrected version.

XML technology also facilitates consulting the dictionary. The project is aiming at an 'à la carte' consultation, where the user will be able to choose the micro-structural elements to be displayed according to his/her competence and needs. The present functionality enabling the user to choose and switch between target languages during the consultation is the prototype of this future system.

The all-XML framework allows a modular and extensible approach illustrated, for example, by the declension and conjugation tables. They are generated from information contained within the XML files by using stylesheets established according to Luxembourgish rules of inflection.

References

Wörterbuch der luxemburgischen Mundart (1906). Lëtzebuerg: Huss.
http://www.lod.lu/.
A Theoretically Computational Lexicon for Turkish Verbs

Elif Eyigoz
University of Rochester
eyigoz@cs.rochester.edu

In this paper, we report progress towards building the first Turkish subcategorization lexicon for verbs, a bilingual Turkish-English electronic dictionary. This preliminary project is based on the proposal by Oflazer et al. (1995), which proposes to represent subcategorization frames as case-frames, accounting for the fact that Turkish noun phrases are overtly case marked. As for semantics, verb senses are mapped to English WordNet (Miller, 1990) definitions. Semantic roles of the arguments associated with the case-frame are compiled from the VerbNet (Kipper, 2000) and FrameNet (Filmore, 1998) databases. The design is currently exemplified for 333 lexical entries.

Turkish is a very creative language in terms of idiomatic compounds headed by verbs (e.g. to give word to someone (to promise someone), to give hand to someone (to support someone), to climb to the head of someone (to abuse someone)). A design for the representation of the compound specific morpho-syntactic properties for Turkish idiomatic compounds headed by verbs was proposed in Eyigoz (2008). This design is also incorporated in the structure of the lexicon.

The lexicon is designed with NLP applications in mind: coding conventions closely follow the conventions of the Turkish Treebank (Oflazer, 2003). The lexicon is compiled in a spreadsheet file, which can be loaded in a database and can be queried via various applications. Data is represented as attribute-value pairs, which is an open-ended data structure that allows for future extension without modifying existing data. The attribute-value pairs and lists are designed in the Prolog style: with brackets and separated with commas. An attribute-value pair is a list of length two, the head of which can be one of the constants in $C = \{\text{lexical, possessor, morph, classifier, is-a, determiner, has-a, modifier, index, occurrence, sentence}\}$. The value of the attribute value pair depends on the attribute in the head position, as exemplified in (1).

(1) \([\text{lexical, elma}]\).

In \([\text{lexical, elma}]\), the value of the attribute “lexical” is the word \textit{elma} (apple). Alternatively, an attribute can be a category in the Turkish Treebank as exemplified in (2).

(2) \([\text{prop, icin}]\).

The value of the pair in (2) is the proposition \textit{icin} (for). Following the coding conventions of the Turkish Treebank allows us to express a wide range of morpho-syntactic restrictions on the elements of the subcategorization frame, and the verb.

An attribute can also be the name of a VerbNet or FrameNet frame, the value of which is the (thematic) role of an argument in that frame. For example, the role of the argument is \textit{victim} in the \textit{abusing} frame in (3).

(3) \([\text{abusing, victim}]\).

(4) \([\text{elifeyigoz, [classifier, [lexical, syntax]], [not[morph,plur]]}]\).
Lists of lists are also used as in (4). The heads of lists can be constants from $C$ (e.g. \textit{classifier} in (4)), or the name of a lexicographer (e.g. \textit{elifeyigoz} in (4)). The members of the tails of lists are also lists of attribute value pairs, or other lists. There are also two prefixes: \textit{not} and \textit{op} (optional). Please note that the prefixes on the lists do not follow the Prolog conventions.

The attribute value pairs and lists are mainly useful in coding the restrictions on the subcategorization frame. However, the restrictions involving the extended syntactic structure, such as control, and the restrictions on the verb, such as voice, are represented in separate columns in the spreadsheet file. For example, the allowable voice markers on the verb (reflexive, reciprocal, causative, passive) are compiled in columns that have values “yes” or “no” for every entry. The justification for this rests on the fact that voice markers may change the meaning of a verb in a non-transparent way derivationally creating a new sense, a sense that we want to disassociate from the entry in question. Currently the lexicon has eight such columns for control and voice markers, and this number will increase with the addition of nominalization properties.

Besides the broad practical needs the lexicon is designed to serve, it is also of great interest to the theoretical linguistic community, as it places more emphasis to depth of lexical representations, thereby drawing a more precise picture of the mental lexicon. Since it contains detailed linguistic information about the morpho-syntactic characteristics of each lexical item, it is a highly sophisticated bilingual electronic dictionary, primarily designed to satisfy theoretical linguistic interests, as opposed to the demands of current NLP applications. The emphasis on the depth, as opposed to breadth of the entries prevents the lexicon to have a wide coverage in the near future, because the foreseeable work on the lexicon involves further enhancing the depth of the representations, but not the coverage.

References


Creating and Exploiting RSS-Based Web Corpora with GlossaNet

Cédrick Fairon, Kévin Macé & Hubert Naets
UCL, CENTAL
Cedrick.fairon@uclouvain.be, Kevin.mace@uclouvain.be, Hubert.naets@uclouvain.be

Lexicographers, Corpus linguists and NLP specialists have been using the Web as corpus (Kilgarriff & Grefenstette, 2003) in mainly two directions that complement each other. A first approach considers the Web itself as a very large corpus. Systems that implement this approach usually offer an interface for querying and concordancing the Web. Basically, they add a layer to traditional search engines like Google or Microsoft Live Search and offer various display options that are useful for linguistic work (concordances, extraction of collocates, stop lists, etc.). Other systems include WebCorp\(^1\) (Renouf, 2003), WebCorpus\(^2\) (Fletcher, 2007), Corpeus\(^3\) (Leturia et al., 2007). A second approach consists in using the Web as a source for extracting texts that will be collected, filtered and recorded in a standalone corpus. Systems of this second category rely on the use of crawlers and filtering techniques. The Wacky\(^4\) project is a typical example of this option (Baroni & Bernardini, 2006) but other examples are numerous.

In this presentation, we will make a demonstration of GlossaNet\(^5\), a free online corpus builder and analyzer. GlossaNet belongs to the second category of tools (it extracts text from the Web and builds corpora) but it is accessible through a Web-based interface, like the tools of the first category. This service integrates two preexisting software: Corporator (Fairon, 2006), a program that creates corpora by downloading and filtering RSS\(^6\) feeds and Unitex, an open source corpus processor (Paumier, 2003). Both software are available as standalone applications.

We will start by demonstrating the publically available interface of GlossaNet. Two steps are involved in using this interface. First, users define a corpus by selecting RSS feeds in a preselected pool of 1500 sources (they can also add their own RSS feeds). A crawler will visit these sources on a regular basis in order to generate a dynamic corpus. Second, the user can register one or more search queries on his/her dynamic corpus. Search queries will be re-applied on the corpus every time it is updated and new concordances will be recorded for the user (results can be emailed, published for the user in a private RSS feed, or they can be viewed online).

Then, we will present a more advanced case study using GlossaNet for generating a dynamic specialized corpus (since RSS are organized by theme, genre or source, it is easy to create specialized corpora) and for configuring a linguistic survey relying on GlossaNet tools and external software. The aim of this survey will be to extract and categorize neologisms and to provide general statistics about word

\(^1\) http://www.webcorp.org.uk
\(^2\) http://webascorpus.org/searchwac.html
\(^3\) http://www.corpeus.org/
\(^4\) http://wacky.sslmit.unibo.it/
\(^5\) Starting on February 1st, 2008, the new version of GlossaNet will be publically available at http://glossa.fltr.ucl.ac.be (but it is currently available for testing on http://130.104.253.24)
\(^6\) RSS is a Web content syndication format: http://www.rssboard.org/rss-specification
frequencies over the time.

References


A Method for Grouping Synonyms

Ingrid Falk1, Claire Gardent2, Evelyne Jacquey3 & Fabienne Venant4
1INRIA & Université Nancy 2, 2CNRS & INRIA Nancy Grand-Est, 3CNRS & ATILF, 4Université Nancy 2 & INRIA Nancy Grand-Est
falk@loria.fr, gardent@loria.fr, jacquey@atilf.fr, venant@loria.fr

Introduction

Because the Princeton WordNet has proved a valuable resource in NLP, many approaches have been developed to support the automatic creation of WordNets for languages other than English. In this paper, we present a method for grouping synonyms and definitions which we believe, can provide the basis for a merge approach to WordNet creation, that is an approach which starts by defining synsets (groups of synonyms) and then structures them into a WordNet type network. More specifically, given a word \( w \), a set \( \text{Syn}_w \) of synonyms for \( w \) and a set of definitions \( \{d_1, ... , d_n\} \) representing the possible meanings of \( w \), the proposed method associates subsets of \( \text{Syn}_w \) with definitions of \( w \). That is, it permits dividing a set of synonyms into subsets and associating each of these subsets with a definition.

Method

Gloss based similarity measures have been used for word sense disambiguation (WSD) (Pedersen, Banerjee & Patwardhan, 2005) to associate a word occurring in a given linguistic context to a given meaning. Similarly, we apply gloss based similarity measures to disambiguate synonyms that is, to choose for each synonym \( \text{Syn}_i \) of a word \( w \), the meaning (represented by a definition \( d_i \) of \( w \)) which fits best.

The definitions used are extracted from a general purpose dictionary for French called the TLFI (Trésor de la Langue Française informatisé). Similarly, the synonyms of a given word \( w \) are extracted from a synonym dictionary for French called Le Petit Robert. For this experiment, we focused on verbs.

To divide the set of synonyms associated by Le Petit Robert synonym dictionary into subsets and associate each resulting synonym subset with a TLFI definition, we proceed as follows:

Index creation. We derived from each TLFI definition an index consisting of the set of lemmatised open class words occurring in the definition.

Computing similarity scores. Given a verb \( v \), \( \text{Syn}_v \) the set of its Petit Robert synonyms and \( D_v \), the set of its TLFI definitions we compute for each pair \(<d_i, \text{syn}_j>\) such that \( d_i \in D_v \) and \( \text{syn}_j \in \text{Syn}_v \) a similarity score using gloss based similarity measures.

Synonym/Definition matching. For each synonym \( \text{syn}_j \) of a verb \( v \) with definitions \( D_v \), we associate \( \text{syn}_j \) with the definition \( d_i \in D_v \) for which the similarity measures give the highest (non null) score.

To assess the impact of the similarity method used, we applied two kinds of similarity measures: word overlap and vector based measures. Word overlap based measures were introduced by Lesk (1986) to perform word sense disambiguation. For a given verb, we compare the index of each of its definitions with the merged indexes of a synonyms definitions. That definition which has the most words in common with
the synonyms definitions is chosen as its most appropriate sense. Vector based similarity measures represent a text as a vector in a word space. Then a TLFI definition can be seen as representing a “direction” in the word space and similarity between words can be computed using some vector similarity measure. The overlap measures we use are simple word overlap, extended word overlap and extended word overlap normalised. The vector based measures used are local word vectors, and second order word vectors with and without a tf*idf cutoff (Pedersen, Banerjee & Patwardhan, 2005).

Evaluation
To evaluate the groupings produced by our method, we built a gold standard consisting of 27 verbs differing in their position (high, medium, low) on three scales (polysemy, genericity and frequency). For each verb, the association between definitions and synonyms was done by 4 professional lexicographers with an interannotator agreement rate of 87%. The resulting reference associates then for each verb in the chosen sample, the set of TLFI definitions and with each definition, a set of synonyms.

To compute recall and precision, we extract the set of triples \(<v, syn_i, def_j>\) defined by the reference such that \(syn_i\) is a synonym of \(v\) which has been assigned to definition \(def_j\) by the annotators. Recall is then the number of correct tuples produced by the system divided by the total number of reference tuples and precision is the same number divided by the total number of tuples produced by the system. F-measure is the harmonic mean of precision and recall. The baseline gives the results obtained when randomly assigning the synonyms of a verb to its definitions.

Results
The 6 similarity measures tried all performed similarly with recall values neighbouring 0.7 (baseline 0.43) and precision 0.75 (baseline 0.44). The best F-measure score (0.71) was achieved with a word overlap measure (extended word overlap normalised).

These results indicate that the TLFI definitions are sufficiently rich to support the application of gloss based similarity measures to the task at hand. That is, even though definitions are typically short, when considering the definitions of two synonyms, the word overlap between two TLFI definitions is sufficiently large to support a meaningful division of the set of synonyms for a given word \(w\) into subsets of synonyms corresponding to the several possible meanings of that word.

Conclusion
The method described here provides a principled way of constructing out of a set of synonyms, a set of synonym subsets each labeled with a (TLFI) definition. Furthermore, it has two features which we believe, make it appropriate as a basis for WordNet construction.

First, it differs from much work on automatic synonym extraction or WordNet construction in that it avoids introducing noise in the data and does not group together words that are not synonyms. This is because it takes as a basis synonym dictionaries. Second, it can be used to merge the content of several synonym dictionaries in a meaningful way. This is because synonym grouping is here mediated by a synonym-to-definition mapping which is independent of the particular synonym groupings listed by synonym lexicons. Indeed, we are currently applying the method described
here to merge five synonym lexicons for French. Further extension of the coverage can be achieved by applying the same method to any available synonym dictionary such as for instance, the Wiktionary\(^1\).

Once sufficient coverage is achieved, the question arises of how the synonym subsets obtained can be linked in a WordNet like structure. We plan to investigate two possible directions. One possibility is to use ontology fusion methods for merging our synsets with either the French EuroWordNet or WOLF (Fiser & Sagot, 2008; Sagot & Fiser, 2008). Another possibility consists in combining our method with a translation approach in order to associate each \(<\text{verb, definition}>\) pair to a Princeton synset. In this way, we can build on the WordNet structure given by PWN and enrich the synsets derived from the five synonym dictionaries with translations of the related english synonyms.

**References**


DIL, an Online Bilingual Specialized Dictionary of Linguistics  
(German-Italian)

Carolina Flinz  
University of Pisa  
carolinaflinz@virgilio.it

This paper describes some selected aspects of DIL (Deutsch-Italienisches Fachwörterbuch der Linguistik). DIL is an online dictionary; bilingual (“monolemmatisch” Wiegand, 1996: 46) and specialized, covering the field of linguistics. The importance of LSP dictionaries and their relevance for the scientific community have been discussed in many works since the 1980s (Wiegand, 1988: 731; Pileegard, 1994: 211; Schaeeder & Bergen Holtz, 1994; Bergen Holtz & Tarp, 1995; Hoffmann, Kalverkämper & Wiegand, 1998).

DIL is still under construction and contains 240 lemmas belonging to the subfield (DaF), but other subfields are in preparation and others are planned. DIL is an open dictionary; participation of experts from the various subfields is welcome. It is now only accessible through the University of Pisa site (http://www.humnet.unipi.it/dott_linggensac/glossword/), but it will soon have its own homepage.

The need for such a tool as DIL was very strongly felt in Italy after the University reform of 1999 (daf 2004: 37). After all there was no written or online dictionary like that. The functions of the dictionary aimed at the intended user group (Storrer & Harriehausen, 1998; Barz, 2005), at their needs, evaluated with the help of a questionnaire (Barz, 2005: 85; Ripfel & Wiegand, 1988: 493) and the situation of use (Schaeeder & Bergen Holtz, 1994; Wiegand, 1977) were crucial for the design and preparation of the dictionary, of its macro- and microstructure.

Other important factors which influenced the web-design, the macro- and microstructure of the dictionary, were the results of two analyses: 1) the analysis of the existing written monolingual dictionaries of linguistics, and 2) the analysis of the existing online dictionaries covering the field of linguistics (both monolingual and bilingual). In particular, the second analysis showed that there is a strong difference between dictionaries in their written and online form; in brief, the written ones are more scientific, they have a more complete macrostructure; the online ones instead do not make use of the most important technological instruments, like search engine links etc. (only 20% of the online dictionaries have a search engine, and only 70% use links). Of course there are also exceptions, but many online dictionaries seem to have been written dictionaries first and later converted into online ones. But online dictionaries are “works in their own right” (Barz, 2005: 103). DIL follows written dictionaries of linguistics as a model (the entries are in alphabetical order; the dictionary has a introduction, an index, abbreviations, a guide for its correct use, guidelines for the addition of new entries, a bibliography) but also tries to use the peculiarities of online dictionaries (Almind, 2005; Abel, 2006: 40). It has search engine, uses links (external to other useful related dictionaries and internal between the entries) and multimedia facilities are also being planned.

In contrast to existing specialized dictionaries within this field, which are essentially monolingual, DIL is a dictionary with many types of information designed
to help the user to produce and translate texts. The dictionary articles contain information about grammar (e.g. genre and plural), synonyms, annotations about the subfield to which a certain term belongs, definitions, examples, cross-references, internet links and an annotated bibliography. DIL can also be used for bibliographical researches.

(1) Example

**Medien, audiovisuelle**

**(Plural)**

**Materiali e strumenti audiovisivi**

**(DaF)** Termine che identifica il materiale e gli strumenti didattici di tipo audiovisivo. Sono caratterizzati da una strumentazione tecnologica e un supporto audiovisivo. Strumenti tecnologici utilizzati 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 sulle attività straniero, presentazione di modelli comportamentali socio-comunicativi adeguati alla situazione, stimolo delle abilità sia ricettive sia produttive (cf)

**Vedi anche:** Video

References


Lexicographical bases for the compilation of the part of the dictionary concerning DaF were a corpus of texts (German and Italian) concerning the subfield of “applied linguistics”. The entries of the existing online written dictionaries of applied linguistics and linguistics were also taken into consideration. Criteria for the selection of lemmas were frequency and relevance (Bergenholtz, 1989: 775). This procedure should be used also for the other sections. The dictionary is a practical tool designed to help both experts in linguistics and “newcomers” to understand, produce texts but also to inform and learn about this field.

In order to make it available to as many users as possible, the dictionary was developed as a free e-dictionary. The publication of selected parts of the electronic dictionary is also being planned.


Acquiring Semantics from Structured Corpora to Enrich an Existing Lexicon

Nuria Gala\textsuperscript{1} & Véronique Rey\textsuperscript{2}
\textsuperscript{1}LIF, Marseille, CNRS-UMR 6166
\textsuperscript{2}SHADYC, Marseille, CNRS & EHESS, UMR 8562
nuria.gala@lif.univ-mrs.fr
veronique.rey@univ-provence.fr

Lexical and semantic information is capital in linguistic resources, be it for language learning purposes or for NLP applications. However, this kind of information is very difficult to collect either manually or automatically. The difficulties come from the nature of the information (what do we mean by ‘semantics’? how is the ‘meaning’ put into words?) and from the resources themselves (how are the ‘semantics’ formalised and displayed?). In this paper we present a methodology to automatically acquire semantics from structured corpora (machine-readable dictionaries and free encyclopaedias on the web). This information is used to enrich an existing lexical database for derivational families of words in French.

Context and problem

In the age of generalized electronic linguistic resources, language learners benefit from a variety of applications. As far as monolingual databases are concerned, for a given unit, the learner obtains etymological, phonological and grammatical information, as well as a definition and several examples illustrating the usages of the word. In addition, the notion of how common a word is, introduced during the nineties by means of frequencies extracted from corpora (Kilgarriff, 1996), is now significantly widespread. The resources used for e-learning are thus a digitalized version 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\textsuperscript{1} (Trésor de la Langue Française) and the DAF\textsuperscript{2} (Dictionnaire de l’Académie 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 on the lexical description (due to storage possibilities, to the conceptual organisation of the information, etc.). Though a number of projects have arisen, especially for multilingual resources (Papillon database\textsuperscript{3}, 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 a same derivational family neither to access a particular lexical unit from a set of ideas carried by the semantic information.

\textsuperscript{1} http://atilf.atilf.fr/tlf.htm
\textsuperscript{2} http://dic.academic.ru/dic.nsf/daf_1835
\textsuperscript{3} http://www.papillon-dictionary.org
Goals and related work

The aim of our project is to enrich with semantics an existing lexicon of French words organized in derivational families. At present, the lexicon, only displays morphological information and is being used by speech therapists for improving the vocabulary learning task of patients presenting particular diseases (dyslexia, Alzheimer). The need of semantic information in this context is twofold: to understand unknown lexical units and access them.

We consider that the learner would better understand the meaning of a lexical unit by grasping the semantic links with other words belonging to the same family (given some unknown words as gluey or glueball s/he would be able to catch their meanings by comparing it to glue, the ‘baseword’, which shares with the former unknown words the notions of sticky, adhesive, viscous). Following (Zock & Schwab, 2008), adding semantics will also allow the learner in another perspective: s/he will be able to find a precise lexical unit from a set of ideas (taking file, key, path, fast s/he will accede to shortcut).

To automatically collect semantic information, we use existing available resources as other authors have already done for different tasks: definitions in machine-readable dictionaries for word sense disambiguation (Lesk, 1986), (Ide & Véronis, 1990); WordNet in an information retrieval task (Richardson et al., 1994); large corpora for building language models (Grefenstette, 2007). In our approach, we combine the information extracted from definitions in available French dictionaries and from Wikipedia.

Experiment and first results

We have started with a preliminary list of words representing ten derivational families: a ‘baseword’ (“terre” – ground- ) and their related terms (“enterrer” – bury-, “territoire” – territory- and “terrasse” – terrace- ). In our approach, we group in a same family French words phonologically close but having, apparently, very dispersed meanings (example, “arme” – weapon- and “alarme” – alarm). The semantic information collected from structured corpora permits to empirically validate the fact that these two words share semantic components (danger, defense, protection).

For each word we have automatically collected dictionary definitions (all of them in the case of multiple meanings -polysemic units and homonyms) as well as the introductory paragraph in Wikipedia (the text before the contents table). After removing stop-words we have gathered counts on frequency and distance (given a target word “enterrement” – burial- we assign a higher weight to “action” and “terre” because they appear closer to the target word than “corps” – body- or “funèbre” – funerary-). The result is a set of lexical vectors with an average of ninety words followed by their distributional information.

*enterrement*

[mettre 1] [cérémonie 0.83] [action 0.81] [rite 0.77] [funérale 0.72] [inhumation 0.5] [terre 1

1 Derivational morphology is very common in French and other Romance languages. In our approach, families of words have an average of ten members.

2 The use of French WordNet was not foreseen in this experience as we needed to collect meaningful lexical units related to a given word, not only synonyms.
We are currently extrapolating this experiment to the whole lexicon (twenty thousand words) and adapting the results to display them in the resource interface. Taking a closer look at lexical vectors provides relevant information to study and measure semantic similarities in a derivational family (Gala et al., 2009).

Word families in Polymots¹ offer a new perspective on the study of words based on phonological stems resulting from language usages instead of traditional lemmas anchored in diachrony. This resource is thus an example of a new approach in e-lexicography offering different possibilities to learn French vocabulary on a basis of phonology and semantics.

References


¹ Polymots will be available on line by end 2009.
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 poster aims to present three editorial orientations designed to improve access to linguistic information at school.

**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:

1. 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;
2. 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.

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. For instance:

- 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;
- 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).
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.

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 explicitly 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 **permettre** when it means **authorize**, contextualisations would illustrate three patterns:

\[
\text{(1) } Qqn \text{ permet } (qqch \ à \ qqn + \ à \ qqn \ de \ V_{\text{inf}} + \ que \ P) : \\
\text{Le maître permet aux élèves les jeux de ballon } \| \text{ le leur permet} \\
\text{Le maître permet aux élèves de jouer au ballon } \| \text{ le leur permet} \\
\text{Le maître permet que les élèves jouent au ballon } \| \text{ le leur permet}
\]

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.

Helping pupils express things in different ways

To enable the dictionary to fully realise its potential as an expression aid, an analysis of the possibilities offered by electronic media has inspired an original way of presenting synonyms and antonyms.

Explicitly linking synonyms and opposites to contextualisations immediately makes clear any adjustments required when substitution takes place. Each contextualisation could have a pull-down menu presenting correlates that can directly replace the item in question, while 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: \(Qqn \ \text{permet à qqn de V_{inf}} / Qqn \ \text{autorise qqn à V_{inf}}\);
- words that are morpho-semantically linked, such as nominal derivatives of verbs in support-verb constructions: \(Qqn \ \text{permet à qqn de V_{inf}} / Qqn \ \text{donne à qqn la permission de V_{inf}}\);
- alternative syntactic constructions: \(Qqn \ \text{permet } (qqch \ à \ qqn + \ à \ qqn \ de \ V_{\text{inf}} + \ Qu \ P)\);
- alternative phrases: \(Il \ \text{est permis à tout le monde de se tromper ! / Tout le monde peut se tromper ! / L’erreur est humaine ! / etc.}\)

Presenting dictionary texts in electronic format offers important perspectives for enriching information and enhancing readability. Though the editorial orientations presented 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


Examined dictionaries

Customising a General EAP Dictionary to Learner Needs

Sylviane Granger & Magali Paquot
Centre for English Corpus Linguistics
Université catholique de Louvain
sylviane.granger@uclouvain.be – magali.paquot@uclouvain.be

The traditional dictionary articles with static data and fixed structures should be replaced 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, 2009: 29)

Vocabulary poses tremendous difficulties for both native English students and learners of English as a Foreign/Second Language (EFL/ESL) in academic settings (Corson, 1997; Evans & Green, 2006). A number of word lists have been compiled to meet the vocabulary needs of university students. The Academic Word List (AWL) (Coxhead, 2000), for example, consists of 570 word families which have wide range and reasonable frequency of occurrence in a large corpus of academic texts, but are not among the 2,000 most frequent English words. Hyland & Tse (2007: 238), however, noted considerable differences in the use of English for Academic Purposes (EAP) vocabulary across disciplines and called into question the very existence of a general EAP 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 students alike, EFL learners have a harder time as they also tend to transfer into English words and phrases that are typical of academic discourse in their L1. This tendency induces a wide range of difficulties: semantic errors (e.g. French learners’ use of ‘indeed in the sense of ‘en effet’), lexico-grammatical errors (e.g. French learners’ use of ‘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’ overuse of 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; Granger & Paquot, 2009b).

In our presentation, we describe a web-based EAP dictionary-cum-writing aid tool, the Louvain EAP Dictionary (LEAD), which takes into account the most recent findings in genre analysis, language teaching and second language acquisition. The dictionary can be customised to learner needs in terms of discipline and mother tongue background. Its underlying principles are close to Tarp’s (2008; 2009) ‘function theory’, which sees lexicographic needs as “related to specific types of users who find themselves in a specific type of social situation which, by definition, is extra-lexicographic” (Tarp, 2009: 25). It 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).

Before using the dictionary, users can select a domain (currently business, medicine or linguistics) and specify their L1 background (currently French). Discipline-oriented customisation is an efficient way to resolve the tension between the particularising trend which advocates the teaching of a more restricted, discipline-based lexical repertoire, and the generalising trend which recognises the existence of a common core EAP vocabulary that can be taught to a wide range of learners in a wide range of disciplines (see Granger & Paquot, 2009a). This is currently being implemented in the selection of examples. The characteristics of good dictionary examples have been clearly identified by Atkins & Rundell (2008: 458): they should be (1) natural and typical, (2) informative, and (3) intelligible. However, these are not intrinsic properties and examples need to be customised to the type of dictionary and the needs of its users. In the LEAD, the noun issue will therefore be illustrated by example (1) when the user has selected medicine as the target discipline and by example (2) when business is the target.

(1) The issue of underreporting of chronic obstructive pulmonary disease (COPD) exacerbations has been addressed by a series of articles, all of which are based on a cohort of patients with COPD living in East London.

(2) A central issue in the debate is whether foreign direct investment produces positive or negative effects of technology spillovers on domestic firms in host countries.

Customisation is also at play in the phraseological treatment of academic words (cf. Granger & Paquot, 2008). In addition to a cross-disciplinary phraseological common core, the LEAD entries also include discipline-specific collocations and lexical bundles, thereby responding 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” (Moon, 2008: 333). Discipline-specific semantic preferences will also be included (e.g. the typical use of the verb analyse in the general meaning of ‘consider something carefully’ in the social sciences, compared to its more specific meaning of ‘use methods to determine the constituent parts or composition of a substance’ in engineering (cf. Hyland & Tse, 2007: 244). As regards access to corpora, a feature that is now regularly included in electronic dictionaries, the LEAD is innovative in giving access to discipline-specific corpora rather than generic corpora.

One of the purposes of L1-background identification is to give feedback on errors and problems that a specific L1 population typically encounters. As well as generic usage notes, different types of L1-specific usage notes are included in the dictionary: error notes, notes on frequency differences between genres and notes on stylistic appropriateness. To this end, we make use of the Varieties of English for Specific Purposes dAtabase (VESPA), a new learner corpus, currently being developed in Louvain in collaboration with several international partners. The corpus contains L2 texts from a wide range of 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).
Another 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 (see Sobkowiak, 2002; Tarp, 2009). The dictionary can be used as a semasiological dictionary (from lexeme to meaning) or 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 & Levitzky-Aviad, 2006) as users who have selected a particular mother tongue background can search lexical entries via their translations into that language.

The Louvain EAP Dictionary is a very flexible tool, which can be customised to the needs of other target populations (e.g. other L1 background populations, other disciplines, other languages). The dictionary is currently a stand-alone product but it could be integrated into a general dictionary and/or a suite of teaching and learning tools.

References

Pajzs (2009). On the possibility of creating multifunctional lexicographical databases.


In this paper we describe the processes for collecting Basque specialized corpora on different domains from the Internet and subsequently extracting specialized lexicons 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.

Our aim is to analyse the usefulness of the Internet as a reliable source of information for specialized lexicography, and to evaluate the contribution to this task of the existing NLP tools for Basque.

**Motivation**

The traditional process for building corpora as we know it is a very laborious and costly one, so corpora built 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 & 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 classic corpora (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 & Grefenstette, 2004):

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

On the other hand, the development of terminological resources is essential for any language that aims at being 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 reasonable cost taking as a data source large samples of real texts (Ahmad & 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 corpora compilation and terminology extraction processes is very attractive indeed.

**Corpora collection**

The compilation of the 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 & Bernardini, 2004).

The system is fed with a sample mini-corpus of documents that covers as
many subareas of the domain as possible (10-20 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 & 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), etc.). A final topic-filtering stage is also added, using the initial sample mini-corpus as a reference and using document similarity techniques (Saralegi & 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%.

**Specialized lexicon extraction**

The lexicon 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 association degree or unithood (Log Likelihood Ratio) (Dunning, 1994). Single terms are ranked according to the termhood or divergence of the word’s frequency with respect to its frequency in a general domain corpus (Relative Frequency Ratio) (Damerau, 1993). Then those candidates that reach a threshold are chosen.

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

**Experiments and evaluation**

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

The domains chosen were Computer Sciences, 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 it is possible that the number 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 the three specialized lexicons. These lexicons were automatically validated against a specialized dictionary, Basic Dictionary of Science and Technology (http://www.zientzia.net/hiztegia/). The terms not appearing in the
dictionary were manually validated by experts.

First we evaluated the domain precision of the lexicons obtained from the Internet, by analysing the distribution of the terms across the domains, taking the distribution in the specialized dictionary as a reference.

Secondly, we evaluated the adequacy of using the Internet as a source. Subcorpora of the aforementioned domains were extracted from a traditional corpus, the ZT Corpus (Areta et al., 2007, http://www.ztcorpusa.net), and lexicons were extracted with the same method used with the web corpora. Then both lexicons were compared.

Finally, the quality of the extracted lexicons was evaluated by comparing them with the lexicons on the domains of a specialized dictionary compiled and recently updated by experts.

The last two steps of the evaluation took three aspects into account:

1. The precision of the extracted lexicons, that is, the percentage of the extracted terms that really belonged to the domain.
2. The recall of the extracted lexicons, in comparison with those of the dictionary.
3. The number of extracted terms that were not in the dictionary lexicons.

The evaluation results are encouraging and indicate that acceptable results can be obtained with much less work than by a completely manual process.

References


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 (Wordnet provides what is referred to as sentence frames for specific lexical units; FrameNet, which describes lexical units within frames, supplies a large number of annotated contexts, along with valency patterns).

The objective of our work is to propose a method for identifying automatically 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 lexical units and their participants in contexts extracted from a French corpus of texts on computing and the Internet.

As shown in Figure 1, the annotation includes semantic information and syntactic data (syntactic function and syntactic groups). The lexical unit annotated is ACCÉDER and the participants identified in this context are [processeur], labeled as an Agent actant, [directement], as a Circumstant of Manner, and [cache], as an Actant of Location. Participants are divided into two groups: namely Actants and Circumstants, the former being necessary to define the lexical unit; the latter being optional (Mel’čuk, 2004).

Although the annotation is based on a methodology developed within the FrameNet project (Ruppenhofer et al., 2006), the roles assigned to participants differ from those given in FrameNet.

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 at reducing the time devoted to manual annotation.

**Automated method for finding relevant participants in contexts**

The manual annotation is a time-consuming task that prompted us to develop an automatic annotation process of participants, more specifically actants. We are convinced that automating part of the process would also provide a valuable assistance to terminologists carrying out this task. This task is divided into three steps: identification of participants, type assignment of participants and semantic roles.
identification.

During the first step, we use the Syntex parser (Bourigault et al., 2005) which computes the dependencies between components of the sentence, as shown in Figure 2.

![Figure 2 Syntex dependencies.](image)

In the results provided by Syntex, we focus on left and right dependencies of the verbal lexical unit (ACCEDE). 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.

![Figure 3 Combination schemata.](image)

The left hand side of the rule (conditions of application) is composed of the sentence’s words with the top links of Figure 3 and the right hand side gives the identified participants. In a rule, a word is described by its features \(<\text{word}_i, \text{Pos}_i, \text{ROLE}_i>\), in which \(\text{Pos}_i\) is the part of syntactic group of word, and \(\text{ROLE}_i\) is its syntactic function. The lexical unit is in uppercase and the result or participant is a word in conditions and is described by the corresponding features. The feature is assigning type actant or circumstant to the participants and semantic roles like Agent, Patient, Destination, etc. The feature \(<\text{Nom}, \text{SUI}>\) or the feature \(<\text{Nom}, \text{OBJ}>\) are Actants in all cases and the feature \(<\text{Adverbe}, \text{ADV}>\) is Circumstant. Here, the type depends only on the features no matter what verbal lexical unit is used. In this case, the corresponding rules are shown in (1), (2) and (3)

\[
\begin{align*}
\text{Lexie} + <\text{word}_i, \text{Nom}, \text{SUI}> & \rightarrow <\text{word}_i, \text{ACT}, \text{Agent}> \\
\text{Lexie} + <\text{word}_i, \text{Nom}, \text{OBJ}> & \rightarrow <\text{word}_i, \text{ACT}, \text{Patient}> \\
\text{Lexie} + <\text{word}_i, \text{Adv}, \text{ADV}> & \rightarrow <\text{word}_i, \text{CIRC}, \text{Manner}> \\
\end{align*}
\]

However for other types of participants, prepositional feature \(<\text{Nom}, \text{NOMPREP}>\) is identified as a participant, see rule (4),

\[
\text{Lexie} + <\text{word}_i, \text{Nom}, \text{NOMPREP}> \rightarrow <\text{word}_i, \text{NOMPREP}>
\]

The type of a participant can also depend on the verbal lexical unit because their complements can be introduced by a preposition and are to be considered as
actants. For example, the complement of **ACCÉDER** is introduced by the preposition *à*.

The sentence in Figure 3 has three participants indicated by the links at the bottom; the application rule to each participant is as shown below in Figure 4.

<table>
<thead>
<tr>
<th><strong>LEXIE+&lt;Processeur, Nom, SUJ&gt;</strong></th>
<th><strong>LEXIE+&lt;directement, Adv, ADV&gt;</strong></th>
<th><strong>LEXIE+&lt;cache, Nom, NOMPREP&gt;</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;Processeur, ACT, Agent&gt;</td>
<td>&lt;directement, CIRC, Manner&gt;</td>
<td>&lt;cache, ACT, Lieu&gt;</td>
</tr>
</tbody>
</table>

**Figure 4** application rules in participants of sentence of Figure 3

When identifying other semantic roles, the roles given by our preliminary experiments are the roles of Agent and Patient associated to features `<Nom, SUJ>` and `<Nom, OBJ>`. For our example, the role of the participant Actant [processeur] whose feature is `<Nom, SUJ>` is Agent and the role of participant Actant [cache] whose feature is `<Nom, NOMPREP>` is Lieu. Other actantial roles are currently being tested.

**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. The application of these rules on more than two thousand manually annotated sentences gave an accuracy of 70% and a recall of 83%. In the future other semantic roles such as Destination, Place or Instrument will be considered.

**References**


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*) which is a bus touring Denmark distributing information on global warming - but only this year. 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). In other words infrequent, often idiosyncratic and highly transient expressions like "osteskuffe" (*cheese drawer*).

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*).
4. new valency of existing words (see example below the table)
5. new multi-word expressions or phrases

when there is evidence that the new word or usage is becoming institutionalized, of course.

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.

Research question: Can machines identify neologisms in natural language text?

The benefits of implementing a semi-automatic neologism extraction system are obvious. Manually monitoring the vast volumes of linguistic usage produced in the modern information society is impossible, and thus manual extraction work is bound to be biased. While assessing the significance of a candidate neologism is a fairly easy task for a trained human (linguist), it is by no means a trivial problem for a computer, however.

As depicted in the poster diagram, we have implemented a semi-automatic
extraction prototype, the "Ordtrawler" (Word trawler). The prototype techniques are very much inspired by the APRIL project (Analysis and Prediction of Innovation in the Lexicon) headed by Antoinette Renouf at the Research and Development Unit for English Studies (RDUES) in Birmingham. Electronic texts are processed by the system as follows:

1. Automatic tokenization and PoS tagging
2. Automatic removal of proper nouns, numerals and non-words (URLs etc.)
3. Optional removal of head words in reference dictionaries (including all their inflected forms)
4. Significance measures
   1. Automatic counting of co-occurring neology markers (e.g. linguistic ones like "so-called" or paralinguistic ones like the use of inverted commas).
   2. Statistical "weirdness" (Ahmad, 1993) (comparing candidate frequencies in the analysis corpus (AC) versus a reference corpus predating the AC).
5. Ranking of remaining candidates by a combined measure of frequency and number of co-occurring neology markers (or "weirdness")

Having completed steps 1 through 3, system output still contains tens of thousands of candidate neologisms (out of an annual input of app. 100 million tokens of newspaper text). Of course, one reason for this vast number of out-of-vocabulary items is that compounds are rendered as continuous strings in the input language. In other words, there can be literally thousands of variants of a simple expression like "0-2-nederlaget" (the 0-2 defeat). Fortunately, applying a simple measure of co-occurring neology markers (step 4) reduced the number of candidates to a much more manageable number (some 2,000).

Interestingly, as is illustrated by the two columns of example candidate neologisms in the table below, the presence of one or more neology markers appear to be a more reliable measure of significance than simple frequency of occurrence (or frequency of the co-occurring markers). For example, "caucus" and "super-duper-tirsdag" are much more frequent than "LAN-party" but they are clearly both ephemeral in nature, and their frequency spike is caused by the presence of many news items referring to the 2008 US elections in the analysis corpus.

This stresses the fact that many neologisms are found among the hapax legomena of the analysis corpus. As was also found by the APRIL project these singletons should not be ignored, because they are statistically significant, and it is an incorrect assumption that they are just typographical errors and ephemera.
<table>
<thead>
<tr>
<th>candidate (status)</th>
<th>markers</th>
<th>frequency (highest)</th>
<th>candidate (status)</th>
<th>markers</th>
<th>frequency (lowest)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;24-års-regel&quot;¹</td>
<td>5</td>
<td>37</td>
<td>&quot;LAN-party&quot;</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>=&gt; include</td>
<td></td>
<td></td>
<td>(LAN party)</td>
<td>=&gt; include</td>
<td></td>
</tr>
<tr>
<td>&quot;caucus&quot; (caucus)</td>
<td>15</td>
<td>28</td>
<td>&quot;billige cigaretter&quot;</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>=&gt; exclude</td>
<td></td>
<td></td>
<td>(cheap cigarettes)</td>
<td>=&gt; exclude</td>
<td></td>
</tr>
<tr>
<td>&quot;super-duper-tirsdag&quot; (Super Duper Tuesday)</td>
<td>9</td>
<td>12</td>
<td>&quot;guldfirer&quot;²</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>=&gt; exclude</td>
<td></td>
<td></td>
<td>=&gt; include</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Our preliminary conclusion on the feasibility of implementing semi-automatic neologism extraction is that it is indeed possible to extract a subset of certain varieties of neologisms with a reasonable precision (in the region of 50%). However, this level of precision could only be attained by using neology markers, and one should not forget that neology based on new valency (e.g. the verb + direct object *dumpe eksamen* ('fail an examination') instead of the former mode of expression with the verb combined with a prepositional group as in (*dumpe til eksamen*) or semantic expansion (e.g. *alumne* ('alumnus') which has been semantically expanded to include the meaning 'member of a society for former students of a university') of existing words cannot be captured by the system outlined in the poster diagram. Automatically detecting these kinds of neology requires sophisticated NLP techniques like parsing, collocational analysis and so on. Techniques which are not always implemented and readily available for minority languages.

Implementing automatic morphological analysis to identify and cluster candidate constructed neologisms (e.g. formed by prefixation or suffixation) would also be a useful, and perhaps technically simpler, augmentation of the system.

**References**


¹ Danish regulations barring family reunification or marriage with foreign nationals for people under 24 years of age
² reference to the Danish boat repeatedly winning the gold medal in Men's lightweight coxless four
Proceedings (pp. 51-76). Heidelberg, Germany.
Elliptical Arguments: A Problem in Relating Meaning to Use

Patrick Hanks
Institute of Formal and Applied Linguistics, Charles University, Prague
patrick.w.hanks@gmail.com

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 and artificial intelligence over the past twenty years, despite many declarations of success, has shown that existing dictionaries, designed for human users, coupled with existing linguistic theory of a top-down, speculative nature, have failed to be suitable for this goal. Nor are results using hierarchical ontologies such as WordNet any better. Such resources are very plausible for human users, but they fail to meet the challenges of mapping meaning systematically onto words in use in ordinary text. In the English-speaking world, vast sums of funding have been poured into computational linguistic research on the ‘Word Sense Disambiguation problem’, using dictionaries for foreign learners such as LDOCE as a resource, with what some of the protagonists (e.g. Ide & Wilks, 2006) now acknowledge are disappointing results. There are several reasons for this, chief among them being the fact that the research question has been formulated on false assumptions. The chief of these is what Fillmore (1976) characterized as a “checklist theory of meaning”. The implication that words have a finite and discrete list of senses that can be disambiguated by some procedure or other is a dangerously crude generalization, encouraged by superficial inspection of traditional dictionaries, but not consistent with the evidence of word use in a large corpus. This paper attempts to show why, and what should be done instead.

Insufficient attention has been paid to the highly variable nature of word usage in ordinary unsupervised text. In this paper, I propose to show that such variations are rule-governed exploitations of norms, which electronic dictionaries of the future will be obliged to make explicit. This implies moving away from ‘Lego-set’ theories of language of a Fregean variety, in which words are put together like children's toy bricks in order to make meaningful propositions. Instead, effective future electronic dictionaries will have to give an account of the normal phraseology associated with each word and how meanings are associated with phraseological patterns, including patterns in which certain arguments are not explicitly realized at all. Thus, a lexical entry word in an e-Dictionary will be no more than an index item, an entry point to an inventory of normal phraseological patterns with which meanings and/or translations and/or other implicatures are associated. This requires a distinction between normal patterns of word use and abnormal uses, which are exploitations of norms.

Although word usage in everyday texts is highly variable, variation is not random. Corpus analysis, using tools such as Sketch Engine (Kilgarriff et al., 2004), shows that, underlying the many variations, usage is highly patterned. Moreover, the variations themselves constitute sets of secondary patterns. A simple example will illustrate the point. Consider the verb fire. Over a dozen different patterns of normal use of this verb can be distinguished by corpus analysis. Some of these patterns activate very similar meanings; others activate quite different meanings. In the most basic pattern, exemplified in (1), the meaning is 'discharge a projectile from a firearm'. This contrasts with other meanings of the same verb, e.g. (2), 'to stimulate or
excite', (3), 'to bake in a kiln', and (4) 'to dismiss from employment'.

In this presentation, I explore 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. Some of these patterns activate similar meanings; others activate quite different meanings. In the most basic pattern (1), the meaning is ‘cause a firearm to discharge a projectile’. This contrasts with other meanings of the same verb, e.g. (2), ‘to stimulate or excite’, (3) to expose to heat in a kiln, and (4) ‘to dismiss from employment’.

(1) I was in a place once when a man fired a gun at me and I did not like it at all.¹

(2) Active citizenship has already fired the imagination of many people.

(3) Fashioning and firing a pot does not affect the clay composition.

(4) This time General Avril fired four lieutenant-colonels.

In these examples, the semantic types of the arguments activate different senses of the verb. In (1), the direct object is a firearm, in (2) it is a psychological entity, in (3) it is a pot, and in (4) it is a person. Each of these direct objects correlates with the semantic types of other arguments; for example, in (1) and (4), there is a correlation with the subject, which is normally a word denoting person, but (1) also correlates with an adverbial of direction ('at me'), which (4) does not. The direct object in (2) typically governs a dependent possessive ('of many people', 'our') and typically correlates with a subject denoting an abstract entity.

It would be very convenient if natural language always behaved in the way suggested by these carefully selected examples of normal patterns of use. However it does not. A pattern dictionary must not only discover and describe phraseological patterns in a way that has not yet been done adequately by any dictionary; it must also be accompanied by a theory (or a rule book) describing the rules that govern variations in each pattern. One such variation is ellipsis (i.e. omission). In ordinary language use, there are some circumstances in which an argument can be omitted, while in other cases they cannot. These omissions rarely bother human readers and hearers, because the speaker or writer correctly judges the omitted item to be 'obvious'. The ellided argument is taken to be common knowledge and therefore not worth stating. Perhaps for this reason, ellipsis has not been adequately described in linguistic theory. However, it can be a very serious problem for computers and language learners alike. It is hard enough to write NLP programs that process data in text; processing data that is not present in the text but that is 'understood' poses an extra challenge. Electronic dictionaries of the future must therefore account for the precise circumstances under which ellipsis is possible.

With (1), both the direct object and the adverbial of direction are optional. One can say:

(1a) I was in a place once when a man fired at me and I did not like it at all

or:

(1b) I was in a place once when a man fired a gun and I did not like it at all.

¹ In this paper, the convention is followed of printing 'real' examples (taken from corpora—BNC unless otherwise stated—and other texts) in roman, while invented examples (used mainly for contrastive purposes) are in italics.
In an appropriate context, one can even say

(1c) *He fired*

and mean he fired a gun. What's more, if the verb is intransitive with no adverbial, the meaning must be that he fired a gun and not that he dismissed someone from employment, nor that he baked a pot in a kiln, nor that his ideas inspired enthusiasm in others. Thus, 1c is quite unambiguous, even though only one argument is explicitly realized. A further complication, as FrameNet shows, is that there is yet another semantic type competing for the direct object slot in 1, namely the projectile. If it is present, it drives out the semantic type [[Fiream]] from this slot, either completely, as in (1d) or into an adverbial slot (as in 1e).

(1d) *A man fired several rounds at me.*

(1e) *A man fired several rounds at me from a revolver.*

By contrast with (1), the patterns illustrated in (2), (3), and (4) do not allow optional omission of arguments. You cannot say 'he fired' and mean that he dismissed somebody. You cannot say, '*Active citizenship fired many people' and mean that it inspired them with enthusiasm. In fact, the latter is ungrammatical as well as invented, hence the asterisk.

If a particular argument type is added to pattern (4), a different sense is activated, or rather an ambiguity arises, requiring additional context to distinguish the meaning.

(4a) *General Avril fired four lieutenant-colonels with enthusiasm.*

(4a) can mean either that General Avril took a sadistic delight in dismissing the lieutenant-colonels or that his ideas inspired them with positive feelings. The wording is identical but the clause structure of each of the two interpretations is quite different. In the first interpretation, the adverbial is an optional adjunct; in the second interpretation, it is a necessary part of the pattern, with an ambiguity resulting from metonymy in the subject: the term *General Avril* (the whole) is being used as metonym for a part or attribute of General Avril (his ideas or policies).

This all sounds very complicated. Does it imply that pattern-based lexicography is impossible? The answer is a resounding no! Pattern-based lexicography is perfectly possible, but firstly we have to reconstruct the theoretical basis of lexicographical practice, and secondly we have to learn to ask the right questions when analysing data. We also have to resist both the temptation to invent examples that fit our preconceived notions about what each word means and the computational linguist's yearning for magic bullets and algorithms that will solve all problems at a single stroke. And we have to be equally sceptical about checklists of word meanings (attached to words in isolation, out of context), as found in almost all current dictionaries. Words in isolation have meaning potential, not meaning, as argued in Hanks (1994, 2000) and elsewhere.

Practical considerations—from the point of view of both lexicographers and dictionary users—suggest that, if a comprehensive account is to be given of the relationship between words and meanings, a two-part structure for lexical analysis of any word is necessary. First, the normal patterns of use must be identified and a robust account given of the meaning of each pattern. Depending on the syntagmatic characteristics of each verb and its position on the cline between idiomaticity and open choice (Sinclair, 1991), this will account for between 70% and 90% of all uses of most verbs. To account for the remaining uses, a set of exploitation rules must be
invoked. One such rule is ellipsis. The dictionary entry for each word must specify precisely the circumstances under which ellipsis—and every other kind of exploitation—is possible.

The presentation concludes by showing sample entries, constructed on the principles just explained, from the Pattern Dictionary of English Verbs (PDEV; http://nlp.fi.muni.cz/projects/ cpal/ -- login as guest via web access). These entries are supported by evidence from BNC and, where appropriate, linked to FrameNet.

References


Verb, noun and adjective complementation is one of the central problems in foreign language production.

English learner’s dictionaries such as the Longman Dictionary of Contemporary English or the Oxford Advanced Learner’s Dictionary contain information in the form of verb patterns (OALD) or pattern indicators of the type want sb to do sth (LDOCE) to address the learners’ productive needs. This kind of pattern information is extremely valuable but the general learner’s dictionaries suffer from two drawbacks: firstly, while they indicate the grammatical structures in which certain words can occur, they are not very systematic as far the lexical aspect of complementation is concerned. Secondly, they address the needs of learners in offering a number of usually common patterns but they do not provide a list of all patterns occurring with a particular word, which, however, is information required by the foreign teacher of English for marking purposes.

Both these needs were addressed in the Valency Dictionary of English (Herbst, Heath, Roe & Götz, 2004), which, however, is not a dictionary intended primarily for learners. The complementation structures identified in this dictionary for English verbs, adjectives and nouns are based on the principles of valency theory. This talk will focus on ways of providing a user-friendly and pattern-oriented way of presenting the information contained in the VDE and outline plans for an online version of the dictionary.

Furthermore, the project of an Erlangen Valency Pattern Bank will be introduced, which is an online database which contains a list of well over 1000 different valency patterns of English verbs. This project can also be seen as an attempt to unite recent constructional approaches to language with long-established item-based ones such as valency theory, which originates at least in part from a foreign language teaching perspective. While both accounts share a distinctly lexical view, Construction Grammar captures generalizations very well whereas the focus of valency theory is to provide detailed information on how to use a certain word in a sentence, including specification of governed prepositions and likely lexical fillings.

It will be demonstrated how patterns can be accessed in a number of ways - i.e. through the complete list of all patterns identified in VDE (sorted alphabetically or according to frequency of occurrence), by searching for particular pattern elements, etc. The patternbank is intended to serve as a research tool for linguists, lexicographers and researchers preparing teaching materials.

References

Berlin & New York: Mouton de Gruyter.
[LDCE]
Automated Collection of Japanese Examples from a Parallel and a Monolingual Corpus

Kristina Hmeljak Sangawa¹, Tomaž Erjavec² & Yoshiko Kawamura³
¹University of Ljubljana, ²Jožef Stefan Institute, ³Tokyo International University
kristina.hmeljak@ff.uni-lj.si, tomaz.erjavec@ijs.si, kawamura@tiu.ac.jp

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; Atkins & Rundell, 2008), are costly to produce (Rychlý et al., 2008). This is especially crucial in the case of voluntary-based or low-budget academic lexicographic projects with limited human and financial resources.

We have undertaken to automate the collection of usage examples for two electronic dictionaries which are being compiled as academic projects with the help of volunteer editors and progressively published on the web: jaSlo (http://nl.ijs.si/jaslo/) - a Japanese-Slovene bilingual dictionary for Slovene learners of Japanese, and Chuta (http://chuta.jp/) - a multilingualized dictionary of Japanese for L2 learners.

Examples for the bilingual dictionary jaSlo (which had ca. 10,000 headwords and only 2370 usage examples) were extracted from a Japanese-Slovene 400,000 word parallel corpus created for this purpose at the University of Ljubljana from available sources: translated lecture handouts (13.5%), teacher-revised student translations (24.5%), literary fiction (15.7%) and multilingual web pages (46.3%). The corpus was automatically lemmatized and aligned, alignment was manually corrected, and all headwords in the dictionary were searched for in the corpus, yielding examples for 4648 lemmas, i.e. approximately half the dictionary entries. Examples were selected according to sentence length (other criteria are planned), appended to the dictionary entries but graphically separated from existing lexicographer-made examples, and published at http://nl.ijs.si/jaslo/cgi/jaslo-eg.pl.

A preliminary analysis of the examples revealed that, although not covering all dictionary entries, the examples proved useful as they yielded some new translational equivalents which had not yet been included in the dictionary, especially less literal meanings, and useful translations of collocations and multi-word expressions. Given the fact that the dictionary was compiled by a small team of contributors with little lexicographic experience, and that it is the first dictionary for this language combination for which hardly any contrastive studies exist yet, it is not surprising that the entries were often incomplete. The automatic collection of examples from a parallel corpus, which was primarily aimed at adding usage examples for the dictionary users, thus actually also proved useful for the dictionary compilers, as a basis for the future enhancement of dictionary entries.

Although there is a vast amount of literature on automatic bilingual terminology extraction from parallel corpora for NLP applications, and despite the
fact that monolingual corpora are nowadays considered indispensable in dictionary making, surprisingly little has been reported on the use of parallel corpora for the compilation of bilingual dictionaries for human readers (Salkie, 2008). In the case of a dictionary project with limited human resources and for a language combination with a shallow research tradition such as ours, the use of parallel corpus data proved to be very fruitful. Further work to be done includes augmenting the corpus itself for better coverage of entries and usages, and enhancement of the example selection procedure to include readability criteria (vocabulary coverage and syntactic complexity, as measured by available Japanese parsers) and typicality criteria (typical collocations and flectional patterns in a reference corpus of Japanese, to make up for possible skewings in the parallel corpus, which is not balanced).

In a similar pilot study, a corpus of usage examples was collected to be combined with Chuta, a multilingualized dictionary for learners of Japanese in which the sense divisions, definitions, usage examples and multi-word units of Japanese entries are first prepared by a team of Japanese speakers, teachers of L2 Japanese, and subsequently translated into different languages (at present 30 languages, including English, Vietnamese, Bulgarian and Slovenian). To balance the sometimes artificial flavour of made-up examples, more examples were collected from JpWaC, a 400 million word lemmatized and PoS tagged corpus of texts collected from the Web (Srđanović et al., 2008).

In the first step we made a 100 million word sample from the JpWaC corpus. All words in this sample corpus were annotated with their difficulty level on the basis of the 4-level vocabulary list in the Japanese language proficiency test specifications (JLPT; Japan Foundation, 2004), which is the most widely used vocabulary list for Japanese as a second/foreign language; words were assigned rankings from level 4 (easiest) to level 1 (most difficult), while words not appearing in the list (including all proper names) were assigned level 0. Sentences were then selected according to empirical criteria, eliminating too short or too long sentences, those with a disproportionate percentage of punctuation marks, numerals, or level 0 words (some level 0 words were retained to account for proper names), sentences with non-Japanese characters (to exclude URLs, English text etc.), with opening and closing quotes or parentheses (because often unintelligible without their wider context), and sentences starting with punctuation or not ending with the Japanese equivalent of a full stop, to eliminate badly-formed sentences and fragments. Five subcorpora were then extracted for each of the five difficulty levels. Subcorpora are composed of single sentences which contain only words of the specified or of easier levels, with the smallest corpus, for the easiest level, containing just under 80,000 words. These corpora are available for Web concordancing at http://nl.ijs.si/jaslo/cqp/. Since both corpus and example collection were automated, relatively little manual labour was required to obtain a sizeable collection of examples, which we plan to link to the multilingual dictionary being developed.

Plans for further work include an enhancement of the selection procedure (applying criteria proposed by Mizuno et al., 2008 and Yoshihashi et al., 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


David Joffe
TshwaneDJe, South Africa
david.joffe@tshwanedje.com

In this presentation, new features of the TLex (also known as TshwaneLex) Dictionary Writing System (DWS) will be demonstrated. The TLex Electronic Dictionary Publishing System, which can be used to publish dictionaries of any type and for any language, will also be demonstrated. It will be shown how TLex has led the way regarding innovative developments in DWS’s, and in many cases has defined what dictionary producers have come to expect and rely on as a baseline set of must-have tools in their ‘eLexicography tool chest’.

Dictionary Writing Systems have evolved substantially over the past ten years (Joffe & de Schryver, 2004). A typical desktop computer today allows all but the very largest dictionary projects to be loaded into memory and edited readily. At the most basic level, the functionality that a standard DWS should offer today includes: deep customizability and flexibility (the ability to create ‘any kind of dictionary, for any language’); a real-time entry preview ‘as you edit’; the ability to work in both stand-alone mode for single-user/offline use as well as networked, multi-user mode for coordinating the work of teams; automatic customizable entry sorting; automatic numbering of homonyms and senses; multimedia support (i.e. the ability to link images and sounds into entries); and real-time data validation checks. Additionally, most users expect tools to be based on open, industry standards such as XML and Unicode, and many prefer true control over their own data and corpora.

Some of the more sophisticated functionality that has also become part of the baseline of a standard DWS include: smart cross-references; a tightly integrated Corpus Query System (CQS); specialized bilingual editing features, such as automated reversal, linked view and “translation equivalent fanouts”; dynamic metalinguage customization; a Ruler Tool for assessing various levels of balance and representativeness across the entire dictionary; ‘multiple dictionaries from one database’; rich statistics; and the ability to create complex search queries.

Part 1 of this presentation will show some of the old features of TLex, with a special focus on the Corpus Query System. The latter is fully integrated into the user’s workflow; merely selecting an entry results in a corpus search being launched for the selected headword in the corpus window. Importantly, this search is done using ‘background processing’, and thus does not even interrupt the user’s workflow – additional matching corpus lines just appear in the corpus window while the user works. A matching line can be ‘grabbed’ as a usage example for the current entry or sense by just clicking on it and activating a keyboard shortcut.

Part 2 of this demonstration will focus on some of the newest features of TLex, some of which may well become standard fixtures in the DWS of tomorrow:

The new “Smart Styles” functionality allows any aspect of the appearance or punctuation of any information in an entry to be dynamically customized in any
imaginable way. As a simple example, the automatic punctuation that appears after a
definition could be dynamically changed, in real time, depending on whether or not a
usage example follows the definition. This system utilises the TLex integrated
'scripting language', allowing unlimited complexity in behaviour.

Another example of “Smart Styles” is the automatic highlighting of the most
frequent headwords based on corpus rank. Most interesting is that this may be done in
real time — if one imports new, updated rankings at any time, the highlighting
immediately updates too.

In principle, it is not a large leap from such adaptive content during dictionary
compilation, to building truly adaptive content into actual electronic dictionaries.

Part 3 of this demonstration will be a brief overview of the TLex Electronic
Dictionary Publishing system. This system allows a publisher to easily publish any
kind of dictionary electronically, for example on CD-ROM. Here again, a
comprehensive feature set defines the baseline functionality that has become the
standard for an electronic dictionary publishing system.

The system integrates with applications such as Microsoft Word, displaying an
instant ‘pop-up’ overlay window showing a miniature version of the dictionary entry
for the word that the user is currently standing on. Other standard features of a typical
electronic dictionary publishing system include: Web integration (e.g. ‘Google search’
and ‘images search’); multimedia support, such as audio recordings of pronunciation
by a mother-tongue speaker; high-speed, pre-indexed searches; customizable search
behaviour using a “search script” to implement language-specific search functionality;
automatic download of “data updates”; complex search queries; and copy-protection
measures.

For bilingual dictionaries, a “Linked View” mode automatically displays
entries related to the current one on the opposite side of the dictionary. It is also
possible to present the user interface in multiple possible languages, while dynamic
meta-language customization allows the user’s language preference to also be applied
to aspects such as the language used for part of speech labels.

Finally, time permitting, the TLex Web Publishing System may also be
demonstrated. The Web has become an increasingly important publishing platform,
despite the challenges faced by publishers in finding working business models. Web
publishing provides opportunities for the introduction of innovative methodologies for
dictionary improvement that historically were previously not feasible, thanks to the
ability to record and analyze actual dictionary usage patterns in detail. The TLex Web
Publishing System gives ‘eLexicographers’ immediate, real-time feedback and
statistics on what information users are actually searching for, allowing new content
improvements to be rapidly focused based on the actual needs of users, even in
response to current news events. This is an implementation of Simultaneous Feedback
(De Schryver, 2010), and for serious dictionary production, should be considered part
of a standard baseline for a Web dictionary publishing system, in addition to other
features such as interface localisation and dynamic meta-language customization.

References

De Schryver, G.-M. (2010). The concept of simultaneous feedback. In R.H. Gouws,
International Encyclopedia of Lexicography. Supplementary Volume: Recent
Developments with Special Focus on Computational Lexicography (Handbooks of Linguistics and Communication Science (HSK) series). Berlin: Walter de Gruyter.
This presentation will describe the principles underlying a new type of Internet open dictionary addressed to students studying lexicography and LSP of tourism with special reference to Florence.

The aim of this reference resource is to provide users with cultural encyclopedic information about local authentic tourist attractions.

The object of the dictionary is Florentine cultural artifacts (names of buildings, monuments, and churches), place names (macro- and microtoponyms) and other proper names, for example outstanding personalities (actors, artists, musicians, public figures, writers, etc.).

This dictionary is structured according to theoretical rules accepted in applied lexicography and includes the following steps:

1. Choosing a city as an object of cultural analysis.
2. Defining the sources of the dictionary (encyclopedias, dictionaries, biographies, guide books, etc.).
4. Defining the entry structure with special reference to its metalanguage.
5. Creating hypertext depending on the compilers’ preferences and knowledge.
6. Adding personal attitude to the entry described.
7. Creating a forum on perfection principles

The planned dictionary has the following mega-, macro- and microstructure.

The megastructure of the planned dictionary contains an introduction and style manual, the dictionary itself and a number of appendices including alphabetic and categorical indices of entry words.

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 microstructure has the following information categories:

1. chronological, regional, statistical labels,
2. definitions with essential encyclopedic information about each entry word,
3. graphic illustrations (pictures, photos, types of prints, sketches, gravures),
4. quotations from literary works, music samples, etc.

The dictionary is characterized by rich multimedia contents, which will be presented in the given way: an entry will have an integrated flash audio and video
player. That feature allows users to watch and listen to the piece of music or video connected with the topic of a certain entry.

It contains audio tracks and videos, interactive flash movies. Such features give users possibilities in obtaining a great amount of information. Due to the Internet format of the dictionary the entry can be expanded to any volume, everything depends only on the scope of information provided for the compilation of the given entry. A user-friendly interface will make the users’ work fast and effective, giving pop-up prompts and additional data in the hints.

This huge hypertext is the advantage of the dictionary in comparison with printed encyclopedic resources and has the following useful links to:

1. Biographical data of the person presented in the entry.
2. Photos of the places he/she lived.
3. Personal attitude of the users’ concerning each entry, etc.

This paper contributes to the theme of the conference as a basis for compiling an electronic dictionary making process which may be used in foreign language teaching and learning within “Cultural Studies Course” and “Course of Cyberlexicography”.

The project may involve along with professional lexicographers, students and any volunteers wishing to participate in compiling of a new type of dictionary for guides and tourists devoted to Florence as a cradle of world culture.

Several sample entries will be demonstrated during presentation.

References

Tickbox Lexicography

Adam Kilgarriff, Vojtěch Kovář & Pavel Rychlíč
Lexical Computing Ltd, UK
Masaryk University, Brno, Czech Republic
adam@lexmasterclass.com; {xkovar3, pary}@fi.muni.cz

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 is then the 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 & Rundell, 2008: 97-101).

In our corpus query tool, the Sketch Engine we have two processes which support the process:

- ‘word sketches’, one-page summaries of the key collocations (and sometimes phrases) for the word, in a table organised by the grammatical relations such as object, modifier, modified etc. (Kilgarriff et al, 2004)
- GDEX, a function for finding good dictionary examples (Kilgarriff et al, 2008)

GDEX opens the way to further supporting the lexicographer by letting them select collocations and examples by ticking boxes rather than by re-typing or using standard cut-and paste. We call this “tickbox lexicography” (TBL).

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
- 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
- they tick a "copy to clipboard" button.

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. (Each target dictionary has its own TBL application). The lexicographer can then paste the structure into the dictionary editing system.

Thus, TBL models and streamlines the process of getting corpus data out of the corpus system and into the dictionary editing system.
We are currently (June 2009) working with a number of dictionary publishing teams on TBL applications for their large-scale dictionary projects. There are two projects where TBL is already in daily use:

1. At the Institute for Dutch Lexicology (INL), the ‘Algemeen Nederlands Woordenboek’ (General Dutch Dictionary, ANW) is a large project running from 2001 to 2019 which has been using the Sketch Engine for corpus access since 2007. In the ANW each example sentence comes with its bibliographic reference. The TBL application gathers author, title, publisher and date and puts them on the clipboard (in an XML structure) alongside the sentence. The dictionary editing software (Niestadt, 2009) has been customised to interpret these XML structures so, when the user pastes from the clipboard into the editor, the different components of the reference are placed in the appropriate database fields.

2. At Macmillan Publishing, the ‘Macmillan Collocations Dictionary’ is in preparation. MCD will start from MEDAL (2007), and will provide a full account of the collocations of the core senses of around 4,000 common and highly ‘collocational’ (Kilgarriff, 2006) words. As in word sketches (and in other collocations dictionaries such as Oxford’s (OCD, 2002, 2009)) collocations will be organised according to the grammatical relations. Some collocations will be illustrated with examples in the paper book; all will 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. 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 lexicography is done, as this will reduce the volume of data to be handled by the clipboard and the dictionary editing system, and will allow the GDEX process to be fine-tuned, using experience during the project as a guide, towards 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 so, 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.

We believe TBL has great potential for both streamlining corpus lexicography and making its outputs more accountable to the corpus.

References


The paper will discuss advantages and disadvantages of the combination of automated information and lexicographically interpreted information in online dictionaries, namely *elexiko*, a hypertext dictionary and a lexical data information system of contemporary German ([http://www.owid.de/elexiko_/index.html](http://www.owid.de/elexiko_/index.html)). It will also contrast *elexiko* to DWDS, a digital dictionary of 20th century German ([http://www.dwds.de](http://www.dwds.de)). Examples of automatically derived information (e.g. automatically extracted citations from the underlying corpus, graphs on paradigmatic relations) and lexicographically compiled information for the same headword in both dictionaries (e.g. paraphrases, information on paradigmatic partners) are provided and evaluated, aiming to develop clear guidelines as to how computerized information and lexicographically interpreted information may be combined profitably in online reference works.

This discussion is necessary because the availability of large electronic corpora has changed the work of lexicographers in more than one way. Nowadays, more and more monolingual and bilingual dictionaries are developed based on exploiting large electronic text corpora by applying corpus-driven and corpus-based approaches (cf. Tognini Bonelli, 2001). How this affected the process of compiling a dictionary has been described for specific dictionary projects (cf. Baugh, Harley & Jellis (1996) or Sinclair (1987)) and has been reflected on in a broader approach in various publications (cf. Klosa (2007), Teubert (1999), Euralex Bibliography of Lexicography: [http://euralex.pbwiki.com/Corpus+Lexicography](http://euralex.pbwiki.com/Corpus+Lexicography), and OBELEX: [http://hypermedia.ids-mannheim.de/pls/lexpublic/bib_en.ansicht](http://hypermedia.ids-mannheim.de/pls/lexpublic/bib_en.ansicht); keyword=corpus-based lexicography).

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 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. Mönich & Kühnberger (2008)) or automatic sense disambiguation (cf. Agirre & 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. Working on a dictionary was exclusively a human task in the past. It is now a combination of applying computer and corpus tools together with the lexicographer’s linguistic competence. In printed dictionaries, this mainly leads to an improvement of 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. Besides the classic inventory of grammatical, morphological, orthographic, semantic, and pragmatic information, Internet dictionaries (e.g. elexiko and DWDS) are able to offer more detailed information and new types of linguistic details.

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). Should automatically retrieved information be trusted, or should its quality be verified by extensive testing before publication? Should the two types of information be marked so that users may be able to rate their reliability? How will users respond to automatically compiled information in a dictionary in general? Will the compilation of an Internet dictionary be accelerated when automated information is included? How does information compiled by lexicographers interact with automated information? Can both be brought together or even linked in a profitable way? Examining elexiko and DWDS with these questions in mind will help to find initial answers to these problems.

References


DWDS - digital dictionary of the German language of the 20th century: http://www.dwds.de/ (08.06.2009)

elexiko: http://www.owid.de/elexiko_/index.html (08.06.2009)

Euralex Bibliography of Lexicography: http://euralex.pbwiki.com/ (08.06.200)

OBELEX – Online Bibliography of Electronic Lexicography: http://hypermedia.ids-mannheim.de/pls/lexpublic/bib_en.ansicht (08.06.200).
One of the goals of a major Slovene lexicographic and Human Language Technology (HLT) project (Communication in Slovene: http://www.slovenscina.eu) which started in June 2008 is the creation of a Slovene lexical database intended both for lexicographic and HLT purposes. Its double nature poses some interesting questions to lexicographers and the HLT community: on one hand, Slovene lexicography would profit from a detailed analysis of Slovene lexica which can later be used for the creation of both monolingual and bilingual dictionaries, and on the other hand, it is necessary to offer a machine-readable database robust enough to be subsequently used for tagging, parsing and semantic annotation of corpora to the HLT community.

The approach taken by the creators of the database combines different aspects of similar lexical database projects for other languages, such as SIMPLE, ELEXICO, STO, ADESSE. However, as the main purpose of the database is lexicographic, the emphasis is on semantic aspects of the lexicon, combined with the syntactic analysis. The Berkeley FrameNet project (Fillmore et al., 2003) was chosen as the main inspiration for lexical data organization. Entry headwords are analyzed on three different levels and five types of data are entered into the database. The first level is the semantic indicator which briefly describes the containing semantic aspect of the entry word use. The second is the argument structure or "semantic frame" which is close to the frame description in the FrameNet terminology. The third level is the syntactic pattern which is mainly but not exclusively intended for HLT purposes. The fourth level is collocational: syntactic patterns are filled with statistically significant collocates and collocations with more restricted patterns are identified. Finally, corpus examples represent the fifth level of data structure.

Phraseology is registered in a separate section of the entry without explicit reference to the sense structure. Phraseological units are entered in their canonical form, together with a semantic indicator and examples of usage.

The procedure involves several steps of analysis. First, a random sample of concordances is exported from the FidaPLUS Slovene reference corpus (Gorjanc et al., 2007) in the Word Sketch Engine (WSE) software (Kilgarriff & Krek, 2006) and analyzed in the Excel file to determine basic sense distribution. Argument structures for different senses are identified on the basis of the concordance sample analysis as well as the word sketch data for the particular headword. Secondly, relevant collocations are exported into the lexical database via TBL (tick box lexicography) feature in the WSE, together with the corpus examples chosen by the GDEX module (Rychly et al, 2008). IDM Dictionary Production System software (Kocjančič et al., 2006) is used for the manual lexicographic work on the database.

The skeleton of the database includes the following information:
Level | Data | Example
--- | --- | ---
I. Lex. entry | Headword | stisniti /to squeeze/
PoS | Glagol /Verb/
II. Semantic | semantic indicator | stisniti z dlanjo /to squeeze with one's hand/
argument structure | ČLOVEK stisne PREDMET z dlanjo
/a PERSON squeezes a THING using his/her hand/
III. Syntactic | syntactic pattern (1) | SBZ1 + stisniti + SBZ4
/NP(nom) + squeeze + NP(acc)/
syntactic pattern (2) | SBZ1 + stisniti + za + SBZ4
/NP(nom) + squeeze + (~by) + NP(acc)/
IV. Colloc. | (1) stisniti [roko, vrat] /to squeeze (one's) [hand, neck]/
(2) stisniti [za ramena, za vrat] /to squeeze (~by the) shoulders, ~by (the) neck/
V. Corpus ex. | (1) Ko je uro pozneje Michaels z landroverjem vozil nazaj h koči, je ob menjalniku otipal ženino roko in jo močno stisnil.
(2) Stisnil jo je za vrat in jo začel daviti ter jo skušal posiliti, vendar mu zaradi motenj s potenco to ni uspelo.
VI. Phraseology | PhrUnit | stisne pri srcu koga ~/sb's heart squeezes/
sem. ind. | občutek tesnobe; skrb
/example | /feeling anxiety, being worried/

Among the main problems identified in the process of sample entry compilation are:

- creating the ontology of semantic types in the argument structure
  (1) SLO. OSEBA rodi OTROKA vs. ŽENSKA rodi OTROKA
  
  (2) Eng. a PERSON bears a CHILD vs. a WOMAN bears a CHILD

- identification and inclusion of obligatory argument structure elements, particularly adverbials of direction, time, location and manner
  (3) Slo. leteti vs. leteti kam
  (4) Eng. to fly vs. to fly somewhere

- automatic extraction of simple vs. "extended" collocations
  (5) Slo. zapolniti mesto vs. zapolniti [prosto, prazno] mesto
(6) Eng. ≈ to fill in a post vs. to fill in an [empty, vacant] post

• differentiation between frozen discursive elements and phraseology, or collocations and compounds

(7) Slo. samo pomisli! vs. pobirati ostanke

(8) Eng. just think! vs. ≈ fig. to pick up the remains

(9) Slo. osnovna šola vs. osnovna izobrazba

(10) Eng. primary school vs. basic education

We will address some of these issues and present the database compilation process.

References


The KXD Shell

Tomer Krisi, Assaf Siani & Ilan Kernerman
K Dictionaries
ilan@kdictionary.com

The K XD Shell has been developed by K Dictionaries over the last couple of years with the idea of creating a generic shell that can be used in varied ways for all our dictionary titles, and is based on our decade-plus experience with developing different electronic dictionary versions and on feedback from several partners and numerous users worldwide. This is a software tool that can absorb dictionary data in any lexicographically-structured XML format and process it into an electronic dictionary application for PC. It can thus also offer a useful solution to other publishing houses who are interested in developing electronic versions of different dictionaries they might have.

KXD has various search options, including advanced and wildcard searches, and a soundex search for words with similar spelling. It is possible to search a single item among the headwords, or to perform a focused search in any of the entry components, including the definitions, examples, expressions, and notes. There is also an 'all-text' search to look up items in the entire dictionary (dictionary-as-corpus).

(1) Example: An All-Text search of the word 'search' in the entire dictionary.

The key features include an index list of all the headwords (and their parts of speech), which can be arranged alphabetically or in any other order. A built-in morphological engine makes it possible to click on any word that appears in the definitions or examples and connect it to its related entry. There is back/forward paging that keeps track of all the entries that were previously loaded, so that any looked-up entry can be reviewed again, yet this history may be erased and restarted at any point. The audio features include human voice pronunciation of the headword and enabling the user to record and listen to their own pronunciation. There are also detailed illustrations with a direct link for each item to its corresponding dictionary entry. The application runs in the background in conjunction with other Windows applications, so the user can look up words in the dictionary while working on another document.

KXD has a double-window display, so when the user clicks on a word in one entry, the related entry of the looked-up word will open in the window underneath or above the first one, thus enabling parallel view of both entries.

(2) Example: Clicking the Portuguese translation ‘nascer’ of the French entry ‘naître’ in the top window opens the Portuguese entry ‘nascer’ in the bottom window.

It is possible to load an unlimited number of dictionaries concurrently, to select the one(s) to work with, and to select the languages for searching a specific item and/or receiving the translation(s). For example, with a multilingual dictionary the user can select any language combination for looking up and viewing results. The KXD shell incorporates a skin engine that makes it possible to change the graphic appearance of the final product, and an activation key that prevents illegal copying. It supports any language, and has a localization kit that makes it possible to change the
Altogether, KXD is a simple yet resourceful electronic dictionary application, based on easy conversion of any XML dictionary data into a lively software product. An initial version was published last year for a Spanish/Norwegian dictionary by Vega in Norway, and this year it is launched with a series of eight new French bilingual dictionaries by Assimil in France, followed by local versions of the semi-bilingual Password English learner’s dictionary, and the 42-language Kernerman English Multilingual Dictionary.
Making a Dictionary without Words

Jette H. Kristoffersen & Thomas Troelsgård
University College Capital, Copenhagen Denmark
jehk@ucc.dk, ttro@ucc.dk

In this paper we will present the new online Danish Sign Language Dictionary. Our focus will be on some of the problems specific to sign language lexicography, especially on the challenges of describing a non-written language and on the advantages and opportunities that new technology offers in this respect.

Until recent times most sign language dictionaries were mere wordlists, with words from a spoken language, each accompanied by a picture of a sign, e.g. Danish words and signs from Danish Sign Language (DSL).

In the late 20th century the scientific exploration of sign language evolved drastically, and a series of new scientifically based sign language dictionaries emerged. One of the main differences in comparison to the old wordlist-dictionaries, has been the wish to treat sign language as any other language, that is to be able to make not only spoken language-sign language dictionaries, but also sign language-spoken language, and of course monolingual sign language dictionaries.

A major challenge in this respect is the fact that no sign language has a written representation that is commonly used by the language's native signers. There does not even exist a common standard of scientific notation, like IPA for spoken languages, although a series of different notation systems have been suggested.

For this reason the phonetic representation of the signs in the Danish Sign Language Dictionary database consists of a series of phonetic features, including, among others, location, handshape and movement. This choice allows us to describe sign pronunciation, but leaves us still facing the problem of deciding how the headwords, that is the signs, should be represented, sorted and searched; a decision which for a sign language dictionary is not obvious, as several approaches are possible. In the Danish Sign Language Dictionary we chose to represent each sign entry by a video, a photo, pictures of the first occurring location and handshape, and a gloss. Glosses are traditionally used for transcribing sign language, and although they can mislead the users by being conceived as equivalents, or as fully covering explanations of the sign's meanings, we didn't find any other means of representation, that was easily read and remembered even by inexperienced users.

Presenting the headwords as videos was considered extremely important, because none of the existing alternatives (both formal notations and photographs with arrows) are as informative as videos, especially regarding showing the hand movement.

Originally it was our intention to publish the dictionary not only digitally, but also as a book, but as it became clear that videos were a main priority, this idea was abandoned in an early stage of the project. The abandoning of the printed book solved another crucial problem: defining the sort order of the entries. Printed sign language dictionaries obviously have to have a fixed order of headwords, for written languages traditionally the alphabetical order, but for sign language dictionaries this is not so obvious, because signs can be sorted by different categories. Printed sign language dictionaries
are traditionally sorted using either location or handshape as the primary key. Our decision of only publishing a digital dictionary made the choice of primary key unnecessary; as the dictionary database includes information about the signs' location and handshape, we could instead simply let the user choose between several possible sort orders, which is a great advantage. Thus, the search result of the dictionary can be sorted using location or handshape as the primary key, along with the possibility of using the "original search result order", where the entries that best match the search criteria are listed first, and where, for text searches, the entries are listed alphabetically according to glosses and equivalents.

Lemmatisation is another major problem facing the sign language lexicographer. Many signs consist not only of a specific configuration and/or movement of the hands, but also of a compulsory accompanying mouth movement or facial expression. Therefore, if sign lemmas are defined solely by manual expression, a lot of signs will be heavily polysemic. On the other hand, mouth movements and facial expressions, are problematic to include as a lemmatisation criterion, partly because they are not as well explored as the manual part, and therefore harder to categorise, partly because they are not always compulsory.

In the DSL Dictionary we chose to include only the signs' manual expression as a lemmatisation criterion. If no further measures were taken, this would result in heavy polysemy, not only because of the missing mouth and face information, but also because of the fact that many DSL signs directly connect to a Danish word and inherit all its meanings, with no regard to their semantic relatedness.

In order to avoid very complex entries, we therefore included semantic relatedness in the lemmatisation criteria. Thus, only closely related meanings and transparent metaphors are placed in the same entry. This decision resulted in more homophones, but also in shorter entries, which composition will, hopefully, be clear, not only to users who know Danish well, but to all users of the dictionary.

References


ABBYY is a world-leading developer of optical character recognition, text processing and language software. It was founded in Moscow in 1989 and back then was primarily focused on electronic dictionaries. The ABBYY Lingvo electronic dictionary was the company’s first product. The first version included an English-Russian dictionary and ran on DOS. Since then, ABBYY Lingvo has considerably evolved, with significant improvements made to each new version. Now ABBYY has a 20-year history of developing software products and conducting research in linguistics, semantics, syntax, and lexicography.

The second part of this demonstration will be devoted to ABBYY Lingvo Content, a dictionary writing system. ABBYY has been developing Lingvo Content since 2003, continuously improving the application based on feedback from ABBYY’s linguists and lexicographers, as well as from dictionary compilers and publishers. So far, ABBYY has released 40 dictionaries created by ABBYY’s lexicographers in collaboration with external compilers. ABBYY has invested its lexicographic expertise into this project, and now almost all of its dictionaries are being compiled with ABBYY Lingvo Content.

Accessing dictionary content with ABBYY Lingvo

ABBYY Lingvo combines reliable content, powerful word look-up capabilities, and intuitive interface in a single software application. It has an integrated morphology engine and a sophisticated full-text search facility. According to ABBYY’s own research, 7 million people use ABBYY Lingvo dictionaries worldwide.

The software supports many usage scenarios, such as translation, quick look-up, and language learning tools. ABBYY Lingvo can be integrated with Microsoft Word, Outlook, and Internet Explorer, from which it can be launched by pressing a combination of hot keys, and with other Windows applications, allowing users to look up unfamiliar words by hovering the mouse cursor over them. ABBYY Lingvo can be used on desktop computers, on mobile devices, and as an online service (with curtailed functionality).

Simultaneous access to various types of content

ABBYY Lingvo aims at integrating monolingual and bilingual content and references, for example encyclopaedias, thematic dictionaries, glossaries, and guides collected from various publishers, including Oxford University Press, HarperCollins Publishers, and other well-known publishing houses. The single-window view enables the user to compare translations from different sources. ABBYY Lingvo provides easy and quick access to lexicographic material, making it a valuable tool for dictionary compilers.
Sophisticated search facility

The full-text search facility, which can be launched automatically, helps users find real usage examples, collocations, or typical word combinations shown in their context. 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 senses of “international” words.

Additional features

ABYY Lingvo’s integrated morphology of about 40 languages offers a range of useful options for end users. For example, users can check inflected word forms, look up words in “non-dictionary” form, or find words even if they are not sure how to spell them. ABBYY’s morphology specialists have also developed a method of translating German composite words into languages that do not have composite forms.

The user can look up a word or its translation in Wikipedia directly from ABBYY Lingvo with a mouse-click. Thus lexicographers can find out how a word is actually used in the media and in online texts.

Usability

Throughout the 20 years of developing ABBYY Lingvo, ABBYY has been paying a lot of attention to the product’s ease of use and graphical interface, striving to cover 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.

Writing dictionaries with ABBYY Lingvo Content

ABYY Lingvo Content allows its users to create dictionaries from scratch, update or supplement existing dictionaries, export 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).

ABYY Lingvo Content is based on client-server architecture. Data in an XML-based format are stored on a database server. Unicode characters are supported (e.g. phonetic transcription symbols and Pinyin are displayed correctly).

Working with entry structure

ABYY Lingvo Content allows changing the structure of a dictionary entry easily by using the “drag & drop” method. For example, the order of senses can be altered in this manner. All entries are well structured and entry sections can be added and combined in strict compliance with Document Type Definition (DTD).

Operations on dictionaries

Several dictionaries can be easily compared. This can be used for updating the word list of an existing dictionary. It is also possible to create sets of entries either manually or by using a sophisticated filter with multiple filtering options whereby users can specify alphabets, content, entry sections, words, labels, entry statuses, and a number of other criteria. For instance, a lexicographer can find all entries that begin with a
certain letter or entries containing a specific word in their usage examples. Filters can also help in managing dictionary projects by assigning selected entries to lexicographers.

**Workflow management and multi-user access**

ABBYY Lingvo Content 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. A history of changes and statistics for each user are also available. Different entry versions can be easily compared and the differences are marked in colour, so that it is easy to see which parts of the entry have been changed, removed or added.

**Exporting dictionaries to ABBYY Lingvo format**

Both the ABBYY Lingvo Content DWS and the ABBYY Lingvo dictionary are useful lexicographic tools, enabling authors to create and update dictionaries and easily export them from the DWS into an end-user application ABBYY Lingvo.

**References**

http://www.abbyy.com/company/
http://www.lingvo.com
Our project in progress is focused 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. The system enables the user to design their own dictionary as they need it, choosing the languages, modifying the entry layout and structure and, if they so wish, use a ready-made description of the source language. Using the system, lexicographers can focus their effort on the contents of the dictionary entirely, as the rest is taken care of by the system. 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.

The main components of the management system are: (1) the EELex system of dictionary administration, (2) an Estonian–X dictionary database, and (3) the management system interface.

1. The EELex system of dictionary administration is a web-based lexicographer's workbench integrating various language technological means: linguistic software and language resources (see Langemets et al., 2006). The main features of EELex are: Unicode support, XML databases, XSD 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 dictionary editing, eg. 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" funded from the National Programme for Estonian Language Technology (2006–2010), see http://www.keeletehnoloogia.ee/projects/leksikograafi-tookeskkond and http://www.keeletehnoloogia.ee/national-programme-for-estonian-language?set_language=en.

At present, the EELex system involves about twenty dictionaries, of which five have been completed (published or finished), eleven are being edited, and three previous dictionaries are being prepared for transferring to an EELex format.

Figure 1 presents an example of the editing window.
2. 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 necessary information – will be supplied by the user compiling the new dictionary.

The EXDD source material comes from a voluminous (ab. 80,000 entries) Estonian–Russian dictionary (1997–2009), plus some material from other dictionaries. The presentation of the material has been tailored for the EXDD, while the (ab. 40,000) entries of a medium-sized dictionary have been subjected to a detailed pre-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 the 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.

Figure 2 presents an example (the entry *kuld*) from the EXDD the way the layout window opens up for the user compiling a bilingual dictionary. Note that the system presents a preliminary “standard” form of the entry, open for the lexicographer to modify if necessary.
Section (1) includes the main part of the entry: headword with a full morphological description <all basic inflectional forms, the inflectional type number, part of speech>. The numbers differentiate between senses with explanations, while each sense is followed by usage examples. Section (2) contains a separate block of the compounds associated with the headword, classified according to the grammatical form and meaning of the first component. Throughout the entry |TE| signals of a translation equivalent to be supplied by the user. In addition, the entry provides structured space for information pertaining the translation equivalent: grammatical information, labels, explanations etc.

Although the EXDD is still being improved and updated, the system is used in-house at present for the compilation of three bilingual dictionaries.

### 3. The dictionary management system for bilingual dictionaries

uses the software of 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 the user 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 (type); 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.

The system involves an Estonian–X dictionary database in an XML format. The user may start their own dictionary either from scratch, with an empty database, or to contract the Institute of the Estonian Language for an EXDD headword list (three sizes available) or the whole EXDD (Estonian half of the dictionary ready made).

Figure 3 presents an example of the window for creating a user's dictionary.

![Dictionary creating window](image)

Figure 3. Dictionary creating window

A public version of the Dictionary management system for bilingual dictionaries will be released as freeware in the autumn of 2009 ([http://exsa.eki.ee](http://exsa.eki.ee)).

**References**


Godelieve Laureys
Ghent University
godelieve.laureys@ugent.be

In this paper I want to demonstrate the results of an experimental Finnish-Danish pre-dictionary version which is derived from two existing bilingual contrastive databases. The process can be described as the semi-automatic extraction of a bilingual database for languages B and C as a result from the linking and merging of two compatible databases viz. for the language pairs ‘A and B’ and ‘A and C’ respectively. I aim to discuss the possibilities of this semi-automatic extraction both from a practical and a theoretical point of view.

In the period 1998-2008 two lexicographical projects leading to the bilingual dictionaries Dutch-Danish and Dutch-Finnish/Finnish-Dutch were successively conducted under my guidance. Within the framework of these projects two electronic contrastive lexical databases were elaborated, which the dictionaries are derived from. These databases contain besides lexical data from the languages concerned also semantic descriptions, grammatical (morphological, categorial and relational) specifications and register-related information (stylistic, geographical and pragmatic) concerning the lexical entries and the examples. These annotated databases surpass the merely lexicographical demands and can be exploited for further (contrastive) research.

Both the Dutch-Danish and the Dutch-Finnish dictionary are based on RBN (Referentiebestand Nederlands). RBN is a corpus-based annotated Dutch lexical database, commissioned by the CLVV and elaborated by a inter-university-consortium under the aegis of the Nederlandse Taalunie. RBN is specifically designed for the making of bilingual dictionaries with Dutch as a source language. RBN 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. Lexical units refer to a specific semantic meaning denoted by a form unit. Bilingual dictionaries are 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. For each LU the database contains a résumé (semantic discriminator) and a short definition, corresponding to the information in a basic monolingual dictionary. RBN also comprises both canonical and contextual example units (EU).

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 which establishes translation-links between lexical units in L1 and L2. OMBI is also equipped with a reversal function on the level of lexical units. This means that the reversing takes place on the level of meaning rather than of form. With regard to polysemous form units this procedure implies that each meaning represented by a
Within the framework of the Dutch-Danish dictionary project quite a lot of time and work was invested in establishing the Dutch macro- and microstructure and making an appropriate selection of the RBN-examples. Not only for reasons of efficiency and cost savings, but also in view of the possibility to create a compatible multilingual database in line with the Hub&Spoke Model, the RBN-version elaborated for the Dutch-Danish dictionary, including the selections, modifications and additions made by our editorial team, was greatly re-used for the Dutch-Finnish dictionary. The adapted version of RBN was re-imported into the editor OMBI 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 micro-structure (i.e. distinction of meanings in polysemous words, grammatical and stylistic annotation and selection of examples).

The result of this working procedures is that we now dispose of an outstanding material allowing us to put into the practice the so-called Hub&Spoke-model developed by Willy Martin. This model pertains to the possibility to attain 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-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 their common link with a given Dutch LU. These links can be identified automatically within OMBI, but need to be processed by advanced queries.

I have tested this model on an experimental basis by merging parts of the Dutch-Danish and the Dutch-Finnish database. By linking Danish and Finnish items (through their univocal and specified relation with a Dutch item) I was able to extract an experimental draft of (parts 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 my lecture I will demonstrate the experiment with some examples, illustrating the benefits of the procedure as well as the gaps that are left.

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 that are best suited for automatic linking and that yield the most operational results. Finally, these simulations may also shed a new light on lexicalisation processes in different languages with regard to words, grammatical and pragmatic collocations and idiomatic expressions.

References

Museum Tusculanum.

Dictionaries

Ease and speed of access to the lexicographic information sought is one important area where electronic dictionaries are expected to excel compared to traditional paper dictionaries (de Schryver, 2003). All too often, users fail to locate information in dictionaries even when the relevant lexicographic data is actually there (Nesi & Haill, 2002). In particular, the task of sense selection in decoding tasks involving (highly) polysemous entries has been found to be problematic for language learners, who will tend to stop at the first sense unless there is a clear indicator that this sense is not appropriate (Tono, 1984).

To remedy the above problem and facilitate quick and accurate access to the relevant sense, a number of English monolingual learners’ dictionaries have recently introduced two types of entry-internal (sublemmatic) access facilitators in longer (=highly polysemous) entries. The first of these are sense indicators given at the beginning of each sense, called signposts, guidewords or shortcuts in the different dictionaries, distributed across senses (Tono, 1997; Bogaards, 1998; Lew & Pajkowska, 2007). The second such access facilitator is the entry menu, a “table of contents” of a dictionary entry, gathering brief sense indicators in a single block in front of the entry proper. The entry menu has been found to improve access speed in paper dictionaries (Tono, 1992).

In our study, we undertake to test the usefulness of entry menus as sublemmatic access facilitators in online bilingual dictionaries, as there have been no studies to date, as far as we know, of how menus (or signposts) work in bilingual dictionaries, and none for electronic dictionaries (though Hulstijn & Atkins (1998) suggest that such a study would be useful). In order to meet this objective, we have created three versions of an experimental electronic bilingual dictionary interface, which will be tested on a sample of approximately 100 intermediate learners of English [we do not disclose nationality to facilitate a fair blind review process]. The three versions, randomly assigned to subjects, are as follows:

1. the complete entry is presented at once (the NO MENU condition)
2. a clickable entry menu is displayed, and upon clicking on a specific sense, the complete entry is displayed, scrolled to the selected sense (the MENU condition)
3. a menu as in 2. above, but in addition the selected sense is highlighted in the full entry (the MENU+HIGHLIGHTING condition)

Subjects will engage in a text translation task while consulting one of the three versions of the experimental dictionary. All of the subjects’ consultation activity will be logged in files, including time stamps. Also, the translation output will be examined and translation accuracy of target words will be scored on that basis. The two dependent variables, duration of lookup and translation accuracy, will be evaluated statistically by ANOVA to determine the overall effect of interface type, and a post-hoc analysis will be conducted to assess the differences between specific conditions.
Thus, the two main research hypotheses are:

1. the presentation of entry menus, rather than whole entries at once, decreases access time to the relevant sense
2. the presentation of entry menus, rather than whole entries at once, improves the accuracy of translation for the sense consulted

In addition, two analogous hypotheses will address the hypothetical added value of sense highlighting over and above the presentation of entry menus.

It is hoped that the results of our study will shed light on the usefulness of menus and sense highlighting in electronic (online in particular) bilingual dictionaries, and will thus provide solid empirical evidence on which to base recommendations for the design of more user-friendly electronic dictionary interfaces with respect to entry navigation.

Our preliminary results indicate that the advantage in terms of access speed comes from sense highlighting rather than entry menu alone. However, when the level of proficiency is factored in, 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 accuracy seems to be largely unaffected by the presence of menus or highlighting.

We will present complete results and recommendations for dictionary-making at the eLEX2009 Conference.

References

Word Frequency Distribution for Electronic English Learner’s Dictionaries: Based on BNC XML

Hanhong Li
Department of Chinese, Translation and Linguistics
City University of Hong Kong
johnlihanhong@yahoo.com.cn

Word frequency information has been an indispensable part of electronic English learner’s dictionaries. The review of the current five major electronic English learner’s dictionaries shows that the word frequency information is mainly based on raw frequency counting with the neglect of word frequency distribution in different text genres. We sum up two approaches as follows:

First, common frequency marking is based on raw counting in large corpora. *Cambridge Advanced Learner’s Dictionary 3rd Edition* (CALD3) divides the word frequency into three categories: 1) essential level which means 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 to make your English sound advanced. *Macmillan English Dictionary 2nd Edition* (MED2) designates three groups of word frequency: very high, high and quite high according to their commonness. *Collins COBUILD Dictionary on CD-ROM 2006* (CCD2006) 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.

Second, refined frequency marking has touched the range problem. In frequency counting, range refers to the number of text categories or samples in which a word occurs (Juilland & Chang-Rodriguez, 1964). Actually, Thorndike and Lorge (1944) had begun to consider the range credit in the vocabulary control movement. It is Kučera and Francis (1967), Hofland and Johansson (1982) who introduced range statistics of fifteen text categories in their wordlist based on modern corpora. Range has been an influential factor for frequency evaluation. *Longman Dictionary of Contemporary English 5th Edition* (LDOCE 5th) takes a further step to mark 3000 most frequent words (named as Longman Communication 3000) in the range of spoken and written English, based on statistical analysis of the Longman Corpus Network. Kilgarriff (1997) has demonstrated that this kind of marking could present information that COBUILD missed. Moreover, to be in the Longman Communication 3000, a word has to be well spread across a wide range of sources (Bullon & Leech, 2007). Though *Oxford Advanced Learner’s Dictionary 7th Edition* (OALD7) does not separate word frequency into different bands, it indicates those key words in a wordlist of Oxford 3000 whose selecting criteria are frequency, range and users’ familiarity.

Although some learner’s dictionaries have considered word distribution in different range of texts when marking the frequency of those important words for English learners, the current electronic learner’s dictionaries never state explicitly 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 is the subcorpus for each text type? Is the
variety of written and spoken texts enough? Are they balanced? Bogaards (2008) challenged that without revealing the proportion of the spoken data used in their corpora the current learner’s dictionaries were mainly based on the written data. His comparison between the data presented in the big-five learner’s dictionaries casts some doubt on the reliability of their frequency indication.

With the further development of modern corpora and studies in text genres and text types, a new vision of word frequency information is available for our dictionary users: word frequency distribution in different genres. Our current research attempts to indicate word frequency information based on our latest categorization of the current text genres in the British National Corpus XML Edition (BNC XML). By reorganizing the current genres in BNC XML (Lee, 2001), we will explore the word frequency distribution across the written genres (academic prose, non-academic prose, literature, news, institutional, religion, pops, business, sports, political law and education) and spoken genres (education, business, institutional, leisure). With the genre tagging in BNC XML, a chart of frequency information across different genres is produced and proposed for our future electronic learner’s dictionaries. With this kind of frequency chart, English learners not only can realize the frequency difference between written and spoken English, but also choose appropriate words for various situations.

References


Electronic dictionaries

Building an OLIF-Based Lexical Database for Representing Constructions

Marc Luder
University of Zurich, Switzerland
m.luder@psychologie.uzh.ch

This paper demonstrates the implementation of a small monolingual lexical database for German (currently 7000 entries) for the purpose of manual and automated lexical queries. The lexicon is part of a web based text analysis application which serves to systematically analyze patients’ 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 interdisciplinary project is located in psychology, but involves likewise linguistics, corpus linguistics, lexicography, interactional linguistics and conversation analysis.

The lexical database is implemented in the OLIF format (Open Lexicon Interchange Format), the XML structure is mapped to a MySQL-database. OLIF 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, Clematide & Distl, 2008). Constructions are single or multi word units, including idioms, metaphors and other phraseological and collocational patterns of different types (Granger & Paquot, 2008; Villavicencio, Bond, Korhonen & McCarthy, 2005), according to the language in the transcripts of 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 for disambiguating 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, such as “Der Kleine Wahrig” (Wahrig, 2007), Duden 11 (Scholze-Stubenrecht & Wermke, 2008), VALBU (Schumacher, Kubczak, Schmidt & de Ruiter, 2004), Dornseiff (Dornseiff, 2004) and FrameNET (Fillmore, Johnson & Petruck, 2003). The lexicon entries include a conceptual category of our narrative analysis coding system. In this regard the main emphasis lies on the actions told and performed by the narrator, i.e. on the coding of verbs and verbal constructions. For this reason we are especially interested in semantic and pragmatic verb classifications of German, as described and proposed by various authors (Čulo, Erk, Pado & Schulte Im Walde, 2008; Fellbaum, Geyken, Herold, Koerner & Neumann, 2006; Fellbaum, 2007; Hanks, Urbschat & Gehweiler, 2006).

The paper presents the structure of lexical multi-word entries and the potential of the analysis in matching the form-meaning pairs of the lexicon against the expressions in the transcribed discourse and disambiguating them (e.g. “mit einem blauen Auge davonkommen”, “to get off with a slap on the wrist”; “ein Zeug

---

1 Erzählanalyse JAKOB: http://www.jakob.uzh.ch
2 http://www.olif.net
machen”, “to make a fuss”). Currently the matching works for lexicalized expressions. We are planning to extend the application by implementing more abstract, only partially lexicalized constructions (e.g. “haben zu [+Infinitiv]”, “to have to [+infinitive]” or “typisch [+human-animate]”, “typically [+human-animate]”).

An intention of our project is to integrate theoretical perspectives of conversation analysis and interactional linguistics with perspectives of corpus linguistics. Speakers use for their utterances a big inventory of prefabricated patterns, from entirely fixed units to more loose phraseological constructs (Moon, 2008). “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 use in our corpus of therapy conversations. A typical way to create a new lexicon entry is as follows: After looking for salient examples of a construction in the therapy transcripts, we look for the constructions in the whole data corpus. Selected quotations are then analysed and interpreted using techniques of conversation analysis (Goodwin & Heritage, 1990). We then search for further quotation examples in alternative corpora (e.g. internet documents) to approve or refuse our hypotheses about the meaning of the construction.

A special issue is the constitution of meaning (Deppermann & Spranz-Fogasy, 2006): Interactional linguists advocate the constitution of meaning as emergent, accordingly meaning is created by speakers locally and in the process of discourse (Günthner, 2007). An investigation of the use of constructions over time, e.g. the analysis of the construction “ein Zeug machen” (“to make a fuss”) over the period of a long-term psychotherapy (300 sessions) shows evidence that the specific meaning of this idiom is more and more fixed over time and that the therapist is adopting the construction from the client (Luder, 2009). This could be an argument for Hoey’s (2005) theory of lexical priming: words and collocation patterns get their individual meaning by frequent use in specific contexts, and 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.

References


Automatic Lexical Acquisition from Corpora, some Limitations and some Tentative Solutions

Cédric Messiant & Thierry Poibeau
Laboratoire d’Informatique de Paris-Nord
{firstname.lastname}@lipn.univ-paris13.fr

Introduction

Large corpora and efficient parsers are now widely available for a growing number of languages. So, even if lexical resources are not always available, it is now possible to acquire large lexicon 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 & Carroll (1997), Korhonen (2002), Schulte im Walde (2002), Messiant (2008a; 2008b). The acquisition process is made of three different steps. 1) first, 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 and 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 if they differ as for their parsing model or filtering strategy.

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 (eg. activity verbs like dormir — to sleep — frequently appear with location complements). The system then assumes that they are arguments, whereas linguistic theory would say without any doubt that these are adjuncts. Lastly, surface cues are sometimes insufficient to recognise ambiguous constructions (cf. …manger une glace à la vanille vs …manger une glace à la terrasse d’un café — to eat a vanilla ice-cream vs to eat an ice-cream outside the cafe).

Some solutions

These issues do not mean that automatic methods are bad, but 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.
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.

The analysis of the co-occurrences of the verb with its complement is meaningful but is not sufficient to fully grasp the problem. 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 & 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.

Lastly, the approach requires a manual validation. Rather than leaving the validation process apart for further (boring) 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. In the end, when too few examples are available, the linguist can provide relevant information to the machine. However, with a well-designed and dynamic validation process, it is possible to obtain accurate and comprehensive lexicons, using only a small fraction of the time necessary when using a fully manual approach.

References

Messiant, C. (2008a). ASSCI: A subcategorization frames acquisition system for
French verbs. In *Proceedings of the Conference of the Association for Computational Linguistics* (ACL, Student Research Workshop) (pp. 55–60), Columbus, Ohio.


LEXDIS, a Tool for Measuring Lexical Proximity

Archibald Michiels
University of Liège, Belgium
amichiels@ulg.ac.be

Collocate lists (alongside with indicators) are among the most valuable metalinguistic information pieces provided by monolingual and bilingual dictionaries to enable the user to distinguish between the various word senses or translations associated with a given lexical item. Most of the time, though, the user’s text does not display as head of the relevant syntactic slot (subject, object, etc.) a morphosyntactic word whose lemmatization yields a member of one of the dictionary’s collocate lists for that syntactic position. The reader (or in our case the computer program) has to go through the items to be found in the collocate lists associated with the relevant syntactic position, measuring the proximity of the textual element to each of the members of every collocate list. The winner (the selected word sense or translation) should be the bearer of the collocate list one of whose members features the best proximity measure with respect to the textual item. For instance, using Defidic (an in-house merge of the Robert-Collins and Oxford-Hachette English-French dictionaries), in

(1) From his beloved sister Fatmeh, who had given him the gilded charm to wear round his neck (John Le Carré, The Little Drummer Girl, Pan Books edition, 1984, p.210)

the analysis of the pair WEAR-CHARM should lead to the selection of the translations porter or mettre (which both display jewellery in their associated collocate lists) rather than arborer, accepter or user whose collocate lists do not provide as good a proximity match with charm as does jewellery. At the same time, charm itself can be disambiguated on the basis of the word sense providing the match, i.e. ci11034 in our CIDE database, namely:

*a small esp gold or silver object worn on a chain as jewellery*

Computing a lexical proximity factor also comes in useful in relating indicators and collocates. Consider 2:

(2) And now that little bogey had been exorcized with the rest! (Angus Wilson, Hemlock and After, Penguin ed., pp. 9-10; French tr. by Marie Tadié (1954), La ciguë et après, Robert Laffont, 10-18, domaine étranger, p. 8: Et voilà que ce petit croquemitaine-là avait été exorcisé comme les autres!)

In order to translate bogey by croquemitaine (in preference to crotte de nez, bogée, épouvantail, démon, spectre and bête noire) we need to explore the collocate list for the object of exorcize (a single list), and attempt to find the adequate reading/translation for bogey by measuring how well the indicators for bogey (namely in nose for crotte de nez, in golf for bogée, frightening for épouvantail/démon, evil spirit for croquemitaine, imagined fear for spectre, and bugbear for bête noire) match against the collocates for the object of exorcize (a single, three-item list: demon, memory, past). The task is simple enough for the human user, but we claim that there is no single lexical resource that can be consulted in order to assess the quality of the match, assessment which proves indispensable for the selection of the right translation, in so far as in the case of the exorcize-bogey pair the indicators and
collocates are the only elements provided by the dictionaries to discriminate between the six candidate translations for *bogey*. LEXDIS, our experimental tool, yields the following proximity ratings (the higher the rating, the closer the lexical items of the pair):

<table>
<thead>
<tr>
<th>Demon</th>
<th>Nose</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demon</td>
<td>Spirit</td>
<td>39</td>
</tr>
<tr>
<td>Demon</td>
<td>Golf</td>
<td>0</td>
</tr>
<tr>
<td>Demon</td>
<td>Frightening</td>
<td>0</td>
</tr>
<tr>
<td>Demon</td>
<td>Fear</td>
<td>0</td>
</tr>
<tr>
<td>Demon</td>
<td>Bugbear</td>
<td>1</td>
</tr>
<tr>
<td>Memory</td>
<td>Nose</td>
<td>0</td>
</tr>
<tr>
<td>Memory</td>
<td>Spirit</td>
<td>6</td>
</tr>
<tr>
<td>Memory</td>
<td>Golf</td>
<td>0</td>
</tr>
<tr>
<td>Memory</td>
<td>Frightening</td>
<td>0</td>
</tr>
<tr>
<td>Memory</td>
<td>Fear</td>
<td>0</td>
</tr>
<tr>
<td>Memory</td>
<td>Bugbear</td>
<td>0</td>
</tr>
<tr>
<td>Past</td>
<td>Nose</td>
<td>0</td>
</tr>
<tr>
<td>Past</td>
<td>Spirit</td>
<td>0</td>
</tr>
<tr>
<td>Past</td>
<td>Golf</td>
<td>0</td>
</tr>
<tr>
<td>Past</td>
<td>Frightening</td>
<td>0</td>
</tr>
<tr>
<td>Past</td>
<td>Fear</td>
<td>0</td>
</tr>
<tr>
<td>Past</td>
<td>Bugbear</td>
<td>0</td>
</tr>
</tbody>
</table>

**BOG(E)Y** is therefore to be translated as:

<table>
<thead>
<tr>
<th>Related through: Indicator <em>spirit</em></th>
<th>45 (39 demon-spirit, 6 memory-spirit)</th>
<th>Croquemitaine selected as translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fear</td>
<td>0</td>
<td>spectre</td>
</tr>
<tr>
<td>Nose</td>
<td>0</td>
<td>crotte de nez</td>
</tr>
<tr>
<td>Golf</td>
<td>0</td>
<td>bogée</td>
</tr>
<tr>
<td>Frightening</td>
<td>0</td>
<td>épouvantail</td>
</tr>
<tr>
<td>Bugbear</td>
<td>1</td>
<td>bête noire</td>
</tr>
</tbody>
</table>

LEXDIS gives a brief protocol of the resources that provide a match and the weights assigned to that match, e.g.

QUERY: [demon, n, spirit, n, w:none, m:g, adjust]
demon with POS=n is related to spirit with POS=n with weight=39 as follows:  
Shared Labels: [my] -> weight: 4  
Shared words in definition: [considered, spirit, supernatural, force, activity, believed, peoples, energy] -> weight: 44  
Cooccurence in Roget's thesaurus -> weight: 1  
Penalty for Heavy Lexical Items -> neg_weight: -10

We claim that lexical proximity cannot be reduced to the measure of relatively well-defined dimensions such as those to be found in a standard WordNet type thesaurus. We argue that it is best conceived of as the result of an inherently heuristics-based exploration of various lexical associations derivable from (suitably massaged) available lexical resources, dictionaries as well as thesauri. The
justification for building such a tool lies wholly in its discriminatory power, i.e. its power to select the right translation or word sense in context.

LEXDIS is designed to measure the proximity of any two English lexical items. It calls on the following lexical resources, available to us mainly through research contracts:

- **semdic**: dictionary clauses derived from CIDE, COBUILD, LDOCE and the WordNet Synsets and Synset Glosses (the dictionary clauses feature the lexical items in both definitions and examples, to the exclusion of a specially designed list of stopwords, both tool words and words specific to lexicographic practice, such as especially)
- **mt**: data base of RC/OH collocates - the pivotal property is co-presence within the same collocate field
- **roget**: database of Roget's Thesaurus Categories (three levels)
- **indic**: data base of RC(Robert/Collins)/OH(Oxford/Hachette) indicators (in these two bilingual E-F/F-E dictionaries, only the E->F direction is explored)
- **coll**: data base of RC/OH collocates - the pivotal element is the collocate bearer
- **envir**: data base of environments derived from RC/OH 'extended' lemmas i.e. including phrases and examples
- **weights**: data base recording the lexical weight of lemmas

LEXDIS is written in SWI-Prolog and runs on standard PCs under the various operating systems for which there exists a SWI-Prolog interpreter/compiler (Windows, Linux, Mac-OS).

**Lexical Resources**

CIDE = Cambridge International Dictionary of English  
COBUILD = Cobuild Dictionary of English, based on the Cobuild Corpus  
LDOCE = Longman Dictionary of Contemporary English  
OH = Oxford/Hachette English-French pair  
RC = Le Robert and Collins English-French pair  
WordNet

**References**

Church, K. & Hanks, P. (1989). Word association norms, mutual information and lexicography. *Association for Computational Linguistics, Vancouver, Canada*  


Digital or electronic lexicography has gained in importance in the last few years. This can be seen by means of the growing list of publications focusing on this field. In the OBELEX-bibliography (http://www.owid.de/obelex/engl), the research contributions from 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 of the Institute for German Language (www.owid.de). OBELEX has been available online and free of charge since December 2008.

OBELEX includes relevant articles, monographs, anthologies, and reviews since 2000 with respect to electronic lexicography, partly also older relevant works. Our special focus is online lexicography. Information on dictionaries is currently not included in OBELEX; the main focus is on metalexicography. However, we are working on a database with 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 are searchable (such as person, title, publication, year). Furthermore, all publications are associated with our keyword list; therefore, a thematic search is possible. The subject language is also noted. For example, one can search for all bibliographic entries from the field of "online lexicography" that deal with online dictionaries in "Spanish" (from a metalexicographic point of view). Altogether with OBELEX, we hope to provide an extensive service for all researchers focused on digital lexicography and dictionary research.
OBELEX: an outline of the main points:

<table>
<thead>
<tr>
<th>Subject area</th>
<th>Articles, monographs, anthologies, and reviews from the field of electronic lexicography with a special focus on online lexicography</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current size</td>
<td>approx. 400 entries</td>
</tr>
<tr>
<td>Systematically evaluated sources</td>
<td><em>International Journal of Lexicography, Lexicographica, Dictionaries, Lexikos, Euralex-Proceedings, proceedings of the International Symposium on Lexicography</em> in Copenhagen as well as relevant monographs and anthologies</td>
</tr>
<tr>
<td>Publication year</td>
<td>Continuously from 2000, partly also older relevant works</td>
</tr>
</tbody>
</table>

With this orientation as far as contents are concerned, the OBELEX-bibliography supplements other bibliographic projects in a useful way: firstly, the printed “Internationale Bibliographie zur germanistischen Lexikographie und Wörterbuchforschung” by Herbert Ernst Wiegand (Wiegand 2006;2007), secondly the "Bibliography of Lexicography" by R.R.K. Hartmann (Hartmann, 2007), and lastly the "International Bibliography of Lexicography" of Euralex (cf. also DeCesaris & Bernal, 2006). OBELEX differs from all these bibliographic projects:

- The Wiegand bibliography is surely the most voluminous, but it has no focal point on electronic lexicography and it will not been carried on. Furthermore, searching for a specific publication is not easy, since the forthcoming register volume has not yet been published (cf. Dziemianko, 2008). Besides this, the main focus is on dictionary research within the field of German Studies. Compared with this, OBELEX includes all international research (but only with respect to digital lexicography).

- Perspective, the Euralex bibliography will include all publications from the Euralex conference proceedings. (The current online version does not include all articles from Euralex conference proceedings.) However, other periodicals or journals are not systematically included. Supplements to this core can be made by a Wiki system. However, OBELEX systematically lists all relevant publications such as contributions from the *International Journal of Lexicography* or the journal *Dictionaries*. Thus, if one is interested in all publications in the Euralex proceedings, the Euralex bibliography is the right choice. If one wants to search in the field of electronic lexicography, the use of OBELEX is advisable.

- The Hartmann bibliography is voluminous and international. However, for anthologies, only the title of the entire book is listed, not the individual articles. Thus, searching for specific articles or reviews is not possible in this bibliography.

As mentioned above, in OBELEX sources have been evaluated systematically since 2000. In the first instance, this limit is set under a pragmatic point of view. Thus, including literature published before 2000, OBELEX wants to fulfil the users’ requirements for finding all relevant literature on the subject electronic lexicography,
for instance, the relevant volumes of the Handbooks of Linguistics and Communication 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. There are mainly monographs from the field of electronic lexicography and articles from journals that are not systematically evaluated. Reviews are also integrated because they often include interesting metalexicographic aspects concerning 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. They should serve as a first insight into the article and may help to find appropriate literature.

In the future, current circulations of the sources will be examined systematically in order to continuously enlarge OBELEX.

It would be very interesting for us to present OBELEX as a poster at eLexicography, especially because the conference visitors are the users we have in mind while compiling OBELEX. So we hope to discuss many features of OBELEX with interested researchers, particularly with regard to the keyword list, the included publications or the search options, and thus receive many suggestions for improving OBELEX.

References


Searching the ANW Dictionary

Jan Niestadt, Carole Tiberius & Fons Moerdijk
Institute for Dutch Lexicology (INL)
The Netherlands
{jan.niestadt,carole.tiberius,fons.moerdijk@inl.nl}

This paper describes the query language we are developing for the ‘Algemeen Nederlands Woordenboek’ (General Dutch Dictionary), further abbreviated as ANW.

The ANW is an online scholarly dictionary of contemporary standard Dutch. The project runs till 2019, but the first results will be published on the web in 2009. The ANW contains a rich information structure including a twofold meaning description – definitions are accompanied by semagrams, a structured representation of knowledge associated with a word – as well as a range of lexical and semantic relations between words (Moerdijk, 2008).

The development of the search application for the ANW is guided by the following requirements. First, a variety of search types needs to be supported including semasiological as well as onomasiological queries. Second, the search types should be modelled after the user’s thought processes. Finally, we would like to be flexible in offering new search types. For these reasons a modular approach was chosen.

In simple search applications, the user interface and search engine are usually strongly coupled: all “search intelligence” (like “apply stemming, search dictionary field A, and if you find no matches there, try searching field B”) is built into the search engine, tailored to exactly the search types you want to offer in the user interface.

As a dictionary becomes more complex, and the user interface no longer closely matches the internal structure of the dictionary, this becomes a problem. A strongly coupled search engine is less generic and thus less reusable, and a great deal of technical expertise is required to adjust search strategies or the user interface.

Commercially available search engines have addressed such problems with various approaches, often by offering a query language. For Apache Lucene1 however (which we use for various reasons, not least among which is familiarity), no well-known, fully-featured solution is available.

Therefore, we are developing a powerful but user-friendly query language which evaluates expressions to yield Lucene query objects. We call it FunQy, which stands for “functional query language”. It is a purely functional language: that is, statements have no side-effects, functions are first-class objects in the language and function manipulation such as composition and partial application are important features.

Here is an example query in FunQy, a simple search for a word form:

\[
\text{WordFormFields} \sim \text{userInput} |? \text{WordFormFields} \sim\sim \text{userInput};
\]

\text{WordFormFields} would be a collection of indexed fields we would like to search

\[
1 \quad \text{http://lucene.apache.org/}
\]
through for word forms. The ‘~’ (‘contains’) operator tests for occurrence of a word. The ‘~~’ (‘contains word similar to’) operator performs a fuzzy search for a word. The logical ‘|?’ (‘or else’) operator performs the right-hand search only if the left-hand search produces no matches. In effect, this query states ‘look for an exact match for word, and look for an approximate match if an exact match is not found’. The result of evaluating this expression would be a Lucene query object which, when executed, hopefully produces the word(s) the user is looking for.

Here is another example, a simple onomasiological search:

\[
\text{OnomasioFields} \sim \text{reduce(filterOut(stopWord, terms(userInput))), or}
\]

‘OnomasioFields’ would be a collection of indexed fields we would like to search through. ‘terms’ splits the user input into search terms. ‘filterOut’ filters out stop words (by specifying ‘stopWord’ as the filter function) and ‘reduce’ creates a ‘pattern set’ from the remaining words by applying an operator (in this example ‘or’) to them. So we search for articles containing any of the words the user typed (minus stop words) in any of the fields we specified. Of course, after this query is executed, some form of ranking needs to be applied to present the best results first. This is outside the scope of FunQy for now, but such functionality may be added in the future.

Using FunQy, we can control almost all features of our search engine. Each search type in our user interface corresponds to a function definition that resides in a FunQy script which the search application executes at startup. This startup script can quickly be modified to experiment with different search strategies.

An interesting side effect of this approach is that we can easily give interested users access to the full range of search features, even if they are not directly available from within our user interface. Users might design their own searches, save them for future use, or even offer them on their own pages.

We hope our modular approach will improve reusability of our search engine and allow people who are not seasoned programmers to experiment with all the features the search engine has to offer, resulting in interesting new search types and better search results for the ANW dictionary.

References

Ontologies in the Mediostructure of LSP eDictionaries

Claudia Oliveira & Pam Peters
Centre for Language Sciences, Macquarie University, Sydney
Claudia.Oliveira@ling.mq.edu.au, Pam.peters@ling.mq.edu.au

Mediostructure has been the missing structural concept in lexicography, only recently delineated by Nielsen (2003). By his account, the mediostructure is the cross-referencing relational structure in a dictionary, i.e. the set of relations existing between data items in the dictionary (headwords, pieces of text in the article, attribute values, user instructions, etc), including both dictionary-internal relations and dictionary-external relations. The mediostructure is a network of connections that should provide chains of information to the user. However, its implications for e-lexicography remain unexplored – which is the focus of this paper.

In an online dictionary one will find cross-referencing (pointers) to use-related contents. These are typically links designed to facilitate the use of web-based application, such as Help, About, FAQ, etc; links to traditional print dictionary features, such as the key to the phonetic transcription symbols; links to instructions on how to use the e-dictionary search system, and to the syntax of the query language. On the other hand, the function-related mediostructure provides the user with a mechanism for chaining together information related to his/her goals. According to Nielsen (2003: 277), the function-related mediostructure should aim to satisfy the particular function of the dictionary. In the case of Language for Specific Purposes (LSP) dictionaries this function is to present the vocabulary of a field of specialization, as exemplified for the lexicography of business by Fuertes-Olivera & Arribas-Baño (2008).

The lexicographers’ expectations from the new e-dictionary format include “dreams revolving around accessibility” of the cross-referencing structure of the dictionary (de Schryver, 2006). This mediostructure access wish-list reflects the perception that “anything, in any multimedia, that can conceivably be hyperlinked and is useful to be linked, will in future also be connected” (Schryver, 2006: 180). The question is how to decide what is useful to be linked in a principled and meaningful way. Complex semantically motivated linkages can be provided through computational ontologies.

The benefits of using computational ontologies for educational purposes have been recognized in the field of Web-based educational systems (WBES), where ontological structures, such as concept maps, topic maps, and conceptual graphs, are employed in organizing, processing, and visualizing subject domain knowledge (Dicheva & Aroyo, 2002). An ontology here is understood as the specifications of “the concepts, relationships, and other distinctions that are relevant for modeling a domain.” (Gruber, 2008). In WBES, specialized subject ontologies can be used as a semantic backbone for pedagogical contents (Dicheva, 2008), which is a feature of what the author classifies as 3rd generation WBES or Semantic WBES (SWBES).

This paper describes research on the use and design of ontological structures as part of the mediostructure of TermFinder, an online LSP dictionary. The TermFinder project (Peters et al, 2008) is a collaboration between academic specialists/teachers and applied linguists/terminographers in creating specialized
online termbanks for pedagogical purposes. The target user of TermFinder termbanks
is the novice student in specialist subjects, who can readily and inexpensively access
the contents via multimodal resources. In terms of the question “who accesses what
where?”, TermFinder termbanks could be characterised as human-readable, human-
oriented electronic dictionaries, available on the internet.

From the user perspective, a term that is networked in a TermFinder ontology
carries on its page a hierarchical display of the links to related terms, up to an
arbitrary relational distance. The links of the hierarchy are labeled with the
relationship between the terms, so that the user has a summary overview of the
conceptual network that underlies the termbank. The example below, from the
Statistics termbank, is a fragment of the hierarchy displayed on the page for
“population”:

Example (1):

```
sample
  is part of: population (is described by: parameter)
    parameter (is estimated by a: statistic)
      such as: population proportion
      such as: median
      such as: population mean
        is a: mean
      such as: standard deviation
    sample (is described by: statistic)
      statistic (estimates a: parameter)
        such as: sample mean
        is a: mean
        such as: sample proportion
        such as: sample variance
```

This horizontal display should be read in a nested way, starting from sample,
as “sample is part of population, which is described by parameter, such as population
proportion, such as median, such as population mean,” and so on. The taxonomic
relations of hyper/hyponym are labeled with “is a” and “such as”, respectively, and
meronymy by the phrase “is a part of”. Domain-specific relations, such as “estimated
by a”, are displayed on the same line to highlight the predication.

The terms in the Statistics termbank were selected through their relatively high
frequency in a customized subject-specific corpus, and on the advice of the
disciplinary expert. The domain-specific ontologies for each termbank have been
designed after most articles were completed, in an inductive process to interconnect
the existing concepts covered by the terms. Methodologically, this may not be ideal,
from the point of view that conceptual analysis at a language-independent level
should be distinct from the analysis of language units. Temmerman & Kerremans’s
approach to terminological analysis (Termontography, 2003: 4) starts with an
ontological analysis of the domain, in order to derive an initial framework of
categories and inter-categorial relationships in a top-down fashion. This framework
serves as “a template for the manual and semi-automatic extraction of knowledge
from a corpus”. Therefore in Temmerman & Kerremans’s view, ontological analysis
of the field should come first. Although this has not happened so far in the case of
TermFinder, ontological analysis has been used recursively in the revision and further
development of the termbanks, especially in enhancing the coverage of the relevant subdomains. In fact, our ex post facto procedure still means that ontological mediostructures contribute to the macrostructure of the termbank and the integration of terms within it.

The main purpose of the paper is to argue for the multiple values of ontologies in an eDictionary project. With their application in the design of the termbank, their use as a non-verbal means to explain the semantic relations of individual terms, and ability to communicate the overarching conceptual structures of the discipline to the student, they are a vital part of the internal mediostructure of e-lexicography.

References


UNITEX, a Corpus Processing System with Multi-Lingual Linguistic Resources

Sébastien Paumier, Takuya Nakamura & Stavroula Voyatzi
Université Paris-Est Marne-la-Vallée
paumier; nakamura; voyatzi@univ-mlv.fr

Unitex is a collection of programs developed for the analysis of texts in natural languages by using linguistic resources and tools (Paumier, 2006). Unitex is an Open Source software: this means that the sources of the programs are distributed with the software, and that anyone can modify and redistribute them under the LPGL licence. The resources consist of electronic dictionaries, grammars and lexicon-grammar tables. They are the fruit of long-term work on the basis of conventional dictionaries, corpora and introspection (cf. Gross, 1986). Similar resources have been developed for other languages in the context of the RELEX laboratory network.1

The electronic dictionaries specify the simple and compound words of a language together with their lemmas and a set of grammatical (semantic and inflectional) codes. The availability of these resources is a major advantage compared to the usual utilities for pattern searching as the information they contain can be used for searching and matching, thus describing large classes of words using very simple patterns. The dictionaries are presented in the DELA formalism (Courtois & Silberztein, 1990) and have been constructed by teams of linguists for several European and non European languages, i.e. French, English, Greek (modern and ancient), Italian, Spanish, German, Thai, Korean, Russian, Polish, Romanian, Finnish, Norwegian, Portuguese (PT and BZ), etc. The latest Unitex package contains resources for many languages.

The grammars deployed here are representations of linguistic phenomena on the basis of recursive transition networks (RTN): a formalism closely related to finite state automata. Numerous studies (Roques & Schabes, 1997) have shown the adequacy of automata for linguistic processing at all descriptive levels from morphology and syntax to phonetic issues. More specifically, grammars created with Unitex and DELA dictionaries have been applied to different areas of research, such as Named Entities Recognition (NER) and Annotation (Martineau et al., 2007; Martineau et al., to appear), Multiword Units (MWUs) Retrieval and Extraction (Mota et al., 2004; Català & Baptista, 2007), Multiword Units (MWUs) Annotation in corpora (Laporte et al., 2008a; Laporte et al., 2008b), Multiword Units (MWUs) Alignment (Vitas et al., to appear), Information Extraction from aligned texts (Paumier & Dumitriu, 2008; Vitas et al., 2008), segmentation in chunks (Blanc et al., 2007A; Blanc et al., 2007b). These grammars are represented as graphs that the user can easily create and update.

Tables built in the context of lexicon-grammar are matrices describing properties of the elements of a language. Many such tables have been constructed for all simple and complex grammatical categories, e.g. verbs, adverbs, light verb constructions, frozen expressions, as a way of describing their relevant properties (Gross, 1994; Leclère et al., 2004). Experience has shown that every word has a quasi

---

2 http://infolingu.univ-mlv.fr/Relex/Relex.html.
unique behavior, and these tables are a way of representing the grammar of every element in the lexicon, hence the name lexicon-grammar for this linguistic approach. Unitex offers a way to directly construct grammars in the form of graphs from these tables (Laporte & Paumier, 2006). Such grammars have been successfully applied to shallow syntactic parsing and analysis of specialised texts (Nakamura, 2004).

Unitex can be viewed as a tool in which one can put these linguistic resources and use them. Its linguistic characteristics are the ones that have motivated the elaboration of these resources: precision, completeness, and the taking into account of multiword units (MWUs), most notably those which concern the enumeration of compound words. Its technical characteristics are its portability, modularity, the possibility of dealing with languages that use special writing systems (e.g. many Asian languages), and its openness, thanks to its Open Source distribution. The following modules have been implemented or re-implemented in order to work with Unitex (some of them are compatible with other systems):

- Elag, a program of lexical disambiguation using manually constructed grammars;
- A program that displays differences between two concordances of the same text;
- Multiflex, a multi-lingual Unicode-compatible program for automatic inflection of multiword units (MWUs);
- An extractor of collocations that provides useful statistical information, e.g. the z-score;
- A morphological analyser that works during lexical analysis and constructs a morpheme-based automaton of input text;
- XALign, an Open Source text alignment tool developed at the Loria.

In this demo, special attention will be paid to some new features recently implemented on Unitex. Taking Named Entity Recognition (NER) as a case of use, we shall first demonstrate how the construction of a pile of local grammars augmented with diverse new context features and how morphological dictionary graphs allowed the normalization of some types of named entities. Secondly, we shall demonstrate the usefulness of merging a text alignment tool with the corpus processor Unitex in order to facilitate the exploration of parallel corpora by using advanced linguistic queries. Thirdly, we shall present a fundamental change regarding the internal representation and handling of textual data inside Unitex: now Unitex performs the segmentation of surface text forms in at least three levels: token level, character level, and part of character level. This functionality is particularly important for languages like Korean.

References


1 http://led.loria.fr/outils/ALIGN/align.html.
of the Workshop on a Broader Perspective on Multiword Expressions, ACL 2007 (pp. 33-40), Prague, Czech Republic.


Processing Collocations in a Terminological Database Based on a Cross-Disciplinary Study of Scientific Texts

Mojca Pecman, Claudie Juilliard, Natalie Kübler & Alexandra Volanschi
CLILLAC-ARP, Université Paris Diderot - Paris 7
mpecman@eila.univ-paris-diderot.fr

The encoding of collocations in a terminological database can provide useful lexical information on conventionalities of languages for specific purposes. Such resources can be very useful to language learners or even to advanced non-native speakers who need to write articles or other text types within the scope of their discipline. This is specifically the case for the speakers using their second language in various scientific fields, as nowadays scientific literature is increasingly 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 higher level of their studies, 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 from the source language to the target language. 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 this problem.

Pursuing this objective, we discuss in this communication the various problems arising from the processing of collocations in a terminological database. Several extensive in-depth studies on lexicographical issues on collocations may be quoted: Benson et al. (1997), Fontenelle (1994; 1997), Hausmann (1979), Heid & Freilbott (1991), L'Homme (1997; 2007), L'Homme & Meynard (1998), Maniez (2001), Meynard (1997), Pavel (1993), Siepmann (2006), etc. 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.

Our study is based on a project carried out by a group of researchers working on corpus linguistics and LSP. The project involves the students working on their Masters Degree in Language Engineering and Specialised Translation who use a predesigned database for encoding terminological and collocational information through a corpus-based study in a variety of domains, such as earth and planetary sciences, medicine, chemistry, biology, mechanics, informatics, etc. Our ultimate objective is to make this database accessible online in order to facilitate the encoding process and to allow an easy access to encoded information for language users.

Bearing in mind our objective of conceiving a procedure for systematically processing collocations in this database, we will discuss a number of lexicographical problems that arise from such a project.

Namely, we will discuss the importance of encoding bilingual collocations (e.g. hypothèse de travail ↔ working hypothesis), synonymous collocations (e.g. to advance a hypothesis, to formulate a hypothesis, to put forward a hypothesis),
antonymic collocations (e.g., *to consolidate a hypothesis* vs. *to invalidate a hypothesis*, *to refute a hypothesis*) and the transformational capacities of collocations (e.g. *to formulate a hypothesis*, *formulation of a hypothesis*).

We will also analyse the use of abbreviations for encoding colligations (e.g. *due to the presence of* <sth>) and semantically related lexical clusters (cf. the concept of semantic preference) allowing to encode a series of lexical items which combine with the same node (e.g. *administer* <un médicament>; < programme> runs). Issues such as representing semantic prosody (e.g. *to cause* <negative phenomenon: ex. an earthquake, a landslide>) in a bilingual database will be addressed.

Furthermore, a syntactic description of collocations will be studied (ex. vb. + term: *to image microdamage, to detect microdamage*, noun + preposition + term: *accumulation of microdamage, detection of microdamage*, etc.).

Finally, we will discuss the possibility of providing information on the meaning of collocations which is expected to facilitate the presentation and hence the comprehension of the collocational structure of the lexicon. A special attention will be devoted to the necessity for differential processing of terminological collocations belonging to particular scientific fields (ex. [medicine] *to image microdamage, to detect microdamage, to label microdamage, to stain microdamage*; [geology] *to extract ice core, to collect ice core, ice melts, rocks melt*, etc.) as opposed to the collocations which are characteristic of a more general scientific discourse (ex. *to put forward a hypothesis, to invalidate a hypothesis, working hypothesis*, etc.).

References

Mobile Dictionaries: Situations of Use

Tadeusz Piotrowski
Wyzsza Szkola Pedagogiczna, Wroclaw
tad46ster@gmail.com

In the paper I will discuss dictionaries for mobile handheld devices (henceforth called mobile dictionaries), which use Microsoft Windows Mobile, both phones and PDAs, with touch screens. It will cover the following topics: situations of use of mobile dictionaries, applications (free of charge), and dictionary data. The principal aim is to work out a model of a typical situation of use of the mobile dictionary, based on the approach similar to that of Sven Tarp (2008). The model can, hopefully, serve as a set of guidelines for the development of mobile dictionaries, and as a preliminary conceptual start for further empirical research. In contrast to Tarp, however, who discusses situations of use based solely on theoretical models, disregarding the existing constraints and habits, I will start from what the field offers now, and will try to develop conceptually various features for the nearest future. I shall argue that the situation of use of a mobile dictionary is cognitively and conceptually different from that of a paper one, therefore one is justified, or even forced, to start model-building in this way. In contrast to paper dictionaries, whose physicality is taken for granted, the users and the producers experiment with the mobiles, basing the changes on what they know from the past, and are constrained by technological limitations (cf. Koyama & Takeuchi, 2004; Taylor & Chan, 1994).

The paper is based on my expert knowledge of the dictionaries, empirically on my practical use of them, predominantly in three languages (also between them and various other languages), Polish, English and Russian, and conceptually on my background as a theoretician and practitioner of lexicography. While many of the issues covered will relate to electronic dictionaries in general, I will focus on those features that can turn the mobile dictionary into the most important text-processing device for human users (cf. Kukulska-Hulme, 2007; Clough et al., 2007).

I will discuss predominantly the reception (passive) mode, in accessing e-text, this mode is more typical than the active one (cf. Lew, 2004 and discussions there). I divide situations of use of mobile dictionaries, as usual, into two broad categories: online and offline, and, correspondingly, dictionaries can be used as background service (typically in online use), and as foreground objects (typically in offline use). The offline uses are similar to those of a paper dictionary, and the major difference is that of a different physicality and a different mode of access. The online uses are more interesting, and, I shall argue, can be seen to be more suited to e-text processing and more convenient for the user. I will deal with the following issues (the list that follows is not exhaustive): navigation in the macrostructure (by which I mean: instant access to all available text on the screen, access to the dictionary wordlist, alphabetic and multimodal, selection of the appropriate entry), navigation in the microstructure (with problems arising because the entry structure of a paper dictionary is ill suited to the small screen), ease of switching between dictionaries, stability of interface.

My study will be based on well-known applications and format of dictionary data, such as ereader, Mobipocket, Microsoft Reader, and lesser known, Russian, such as Haali Reader and zddictionary, and Chinese products, such as mdict.
Finally, I will touch on the issue of the user as a dictionary maker: it is interesting that mobile dictionaries quite often allow the users to manipulate dictionary data in ways most suitable and convenient to them. This relates to the format of dictionary data: to what extent the format is open, accessible, and adequately documented, i.e., which format allows the user to efficiently process his or her data.

References

FDVC - Creating a Corpus-Driven Frequency Dictionary of Verb Phrase Constructions

Bálint Sass & Júlia Pajzs
Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary
sass.balint@nytud.hu, pajzs@nytud.hu

We present a method for creating a special dictionary which:
• is a corpus-driven (Tognini-Bonelli, 2001), frequency dictionary;
• has verb phrase constructions (VPCs) as entries;
• is a meaningless dictionary in the sense of Maarten (2008);
• is for Hungarian but the core methodology is language independent;
• is created in a mostly automatic way with less manual lexicographic work;
• can be created with a low budget;
• is hoped to be useful in both language teaching and NLP.

We follow the Sinclairian approach of corpus-driven lexicography. We take a corpus and “jettison ruthlessly” (Hanks, 2008) 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. Knowing that “authenticity alone is not enough: evidence of conventionality is also needed” (Hanks, 2008) we take the most frequent 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 (consisting of a verb plus some NP or PP dependents) from verb subcategorization frames through light verb constructions to completely rigid figures of speech. The so called multiword verbs (e.g. to take sg into account or to get rid of sg) are at the heart of our approach. Having fixed (like the object rid above) and free (like the of position above) positions both they are borderline cases between verb subcategorization frames and real multiword expressions. Contrary to common intuition, they are expressly frequent, they can not be treated marginally. If we take the fixed position as a part of the multiword verb itself, we can treat simple and multiword verbs the same way: both can have some free positions beside. Entries in FDVC are VPCs, the microstructure apparently integrates phraseology as the basic units are phrases. We arrange the VPCs around a verb in a subsequent step to form more traditional dictionary entries.

On the one hand, the FDVC can be called a “meaningless dictionary”. It does not contain explicit definitions, just enumerates the frequent VPCs together with corpus frequencies. Most dictionary users are looking up only basic information, for these tasks meaningless dictionaries are efficient (Maarten, 2008). On the other hand, it contains also a corpus sentence exemplifying the meaning. Furthermore, this meaning is fairly concrete, as VPCs (being collocations) usually have one and only one meaning (Yarowsky, 1993). In fact, most VPCs are real constructions, “form and meaning pairings” (Goldberg, 2006), as they cannot be broken down into smaller units without loss of meaning. Each VPC represent a pattern of use, and can be paired with one sense of its main (simple or multiword) verb.
The dictionary creation process is mostly automatic: starting from the morphosyntactically tagged and disambiguated Hungarian National Corpus (HNC) (Váradi, 2002) we obtain raw dictionary entries using some NLP tools; only the final editing step is manual lexicographic work. We proceed the following way:

HNC →
1. chunking to have verbs and NP/PP dependents;
→ corpus (with richer annotation) →
2. an algorithm based on cumulative frequency of corpus patterns to collect frequent VPCs, with appropriate treatment of fixed and free positions (details and evaluation can be found in Sass (2009)); and another algorithm to collect suitable examples for VPCs;
→ frequent VPCs →
3. manual lexicographic work: error correcting and example selecting
→ dictionary

It should be emphasized that step 2 supersedes a substantial amount of manual lexicographic work. As a result of this step VPCs (arranged around verbs) are presented in XML form, so the lexicographer can edit the entries in an XML editor. In step 3 he/she basically has to check if the patterns suggested by the program are correct, and to choose among the example sentences the most appropriate ones. When doing this, the suggestions made in (Kilgarriff, 2008) are taken into account (choosing full-sentence examples, or at least clauses with full predicate structure, avoid personal names etc.). Sometimes none of the example sentences are correct or appropriate for illustration, in these cases other ones are retrieved from the HNC by a special corpus query system (Sass, 2008). In this form, the task of the lexicographer is considerably easier and the result needs much less corrections than before.

It should be noted that the algorithms in step 2 are language independent, so the methodology can be applied to any language, if we have a POS tagger and a suitable chunker. This methodology allows creating smaller budget dictionaries as the programming and support costs (step 1 and 2) are estimated to 1 man-year, and the lexicographic work (step 3) is also about 1 man-year for a dictionary containing about 3000 verbs and 8000 VPCs altogether.

Beside the traditional (alphabetically ordered by verb) presentation we plan to have several indexes. All of them can be generated automatically:

- aggregated list of all VPCs sorted by frequency - in fact this is the true FDVC;
- an index by general patterns (i.e. VPCs without the verb);
- an index by number of fixed/free positions;
- a frequency list of verb stems;
- an index by lemmas in fixed positions.

Here is an example entry for the verb *elver* (*to beat*) in XML form. It is in the stage after step 2 (amended by manually choosing one corpus example from the auto-
generated ten for each VPC).

The corresponding dictionary entry showing the most important three verb phrases constructions of this verb is:

elver [744]
elver –t [284] hogy minap elvertelek azért, ...
elver jég -t [36] Már ahol a jég nem verte el a termést!
elver -n por-t [95] vagy hogy egy pár túlbuzgó helyi tanácselnökön verjék el a port.

English translation of the entry:

beat [744]
beat OBJECT [284] that I beat you yesterday, because ...
beat ice OBJECT [36] Just where the hail did not destroy the crop!
beat ON dust-OBJECT [95] or to blame some overzealous local mayors.

Verb phrase constructions are translated word by word while example sentences have overall translations, so it can be seen that when hail destroys something Hungarians say the ice beats it; and to blame sy is put in Hungarian something like to beat the dust on sy.

We described the creation of a Corpus-driven Frequency Dictionary of Verb Phrase Constructions (FDVC) for the Hungarian language. We collected automatically all VPCs from corpus, and presented them to the lexicographer in a convenient XML form, significantly reducing the manual lexicographical work this
way. Core algorithms are language independent. Using this methodology we can obtain a lexical database, which is at first a good learner dictionary which lists all frequent VPCs and “helps students to write and speak idiomatically” (Hanks, 2008). Beyond that, it is a rich lexical resource from which many NLP tasks could benefit from information retrieval to machine translation.

References


The Bilexicon: A New Resource for Language Learners

Dirk Siepmann
University of Osnabrück, Germany
dsiepmann@t-online.de

So far little research effort has been expended upon describing the extent of native-like lexical competence in the L2. There is only one study for the language pair German-French (Schmidt, 2005), whose aim it is to list a large section of the receptive vocabulary of French which is ‘intransparent’ from a German perspective.

What Hausmann has achieved for the receptive side the Bilexicon project aims to do for the productive side: to draw up a near-comprehensive list of those collocations (including colligations) which may be considered to make up a native-like vocabulary and to enable students to learn this vocabulary from a learner’s thesaurus.

The compilation of the native-like vocabulary proceeds from two premises:

1. Any attempt to determine basic and advanced vocabularies must start from a list of all native-speaker signs (perhaps even including manual and facial gestures), i.e. the entire lexicon of the language. The approach is thus essentially top-down.

2. It is from such a list that a near-native vocabulary can then be constructed.

Thus, rather than asking, as the traditional frequency approach did, ‘which are the most frequent words in the language, and which words do we need to add to these to obtain a good working vocabulary?’, this approach poses the question ‘what are the meaning units (i.e. collocations) that native speakers use, and which of these have to be mastered to be able to perform at a near-native (or lower) proficiency level?’.

The Bilexicon project stands in the long tradition of what, borrowing from McArthur 1986, we might call ‘thematic learner lexicography’; recent examples of this tradition include the Longman Lexicon of Contemporary English, Vocabulaire anglais et américain (Robert/Collins) and Cambridge Word Routes, to name but a few.

What distinguishes the Bilexicon from other bilingual thesauri is a) that it focuses on collocations and b) that allocation of entries to topic areas is essentially bottom-up, that is, it is the collocations found in the subject-specific corpora which determine the setting up and internal structuring of sub-areas and situation types. This stands in contrast with traditional approaches to thesaurus building, where terms were inserted into a fully pre-determined ontological structure. There are, of course, obvious limitations to such an approach in that some words and collocations have both general and topic-specific uses. A case in point is the vocabulary relating to damage, which is important in such situation types as ‘car accidents’ but may also apply to a wide range of other situations (any kind of accident, intention to harm, legal terminology, etc.).

The dictionary has been implemented on the web. Underlying the thematic organization in the web version will be a layer of semantic links inspired by such work as Francis, Hunston & Manning (1996, 1998) (i.e. meaning groups based on
valency patterns) and Mel’čuk’s lexical functions (i.e. collocational meaning groups). This will enable users to extend their vocabulary along a non-thematic route and will raise their awareness of the close link between sense and syntax.

References

Lexicographers rely upon lexical knowledge bases that provide consensus, integration and clarity of concepts and definitions within specialty fields. One field of distinct complexity, especially in countries like Mexico, is that of sexuality studies and sexual education. Perhaps the main problem for a dictionary for this field is the dispersion of terminology. A dictionary tailored to the needs of the Mexican audience presupposes, on the one hand, some articulated set of concepts and definitions representing the up to date sexological knowledge: the conceptual system deemed by experts to be most relevant in the field of sexology. On the other hand, the set of terms used by regular Mexican people to refer to their sexual experiences should also be considered. Indeed, the sociolinguistic dimension is a very complex issue because regional and social variation is inevitable.

The present paper describes our attempt to deal with the building of a lexical knowledge base for such dictionary. Initially, we describe our criteria to build an electronic corpus for a Mexican Spanish dictionary of sexuality (Medina & Sierra, 2004). This corpus was built with one principal goal: to examine lexical reality about sexuality in Mexico. In that sense, it was not conceived as a one purpose corpus. In other words, we foresee that this corpus will have other uses besides term acquisition and other information extraction tasks. In fact, this resource will be open to the general public. Once built, the corpus is used for terminology extraction of the elementary terms in all fields of sexuality. Then, a rich knowledge base is created, based on definitions given by diverse people and institutions, from experts to laypeople using the terminology. This base will provide lexicographers with wide and representative definitions.

The gathering of representative discourse of the field is the first step towards constructing the electronic corpus. The target size for the corpus is two million tokens; this being an adequate size for obtaining a terminology of 2500-5000 types. This terminology will be obtained with the integration of the tools that are described below. Although we are not working on a complete terminology dictionary, our intention is to find all the terms extractable from any large corpus.

Proper research begins with the selection of the sample texts. First, the relevant thematic areas of the field in question are identified. Second, the key texts of the field are included; in order to get a balanced corpus, we consider five levels of expertise, from the highest one given by experts, to the lowest level given by chats and virtual forums where people use colloquial terminology.

Although we have scanned and OCRed printed documents, we prefer using the Internet to retrieve online documents. We did not use a web crawler or any other automatic system to retrieve the documents that we used. Then we take into account our design criteria, i.e. different level of expertise, national origin of documents and thematic area. We use a part-of-speech tagger (POST) for Spanish which not only gives the grammatical category of words in the text, but also gives their lemmas.
Once the corpus is built, we proceed to extract the terminology by two complementary methods. First, we apply an improved automatic multiword term recognition system for specialised documents (Barrón-Cedeño et al, 2009), in order to extract both single and multiword terms in Spanish texts. Second, we accomplish terminology extraction by comparing the distribution of words from a general purpose corpus with that of a specialised corpus. Selection of those word types with higher frequencies in the latter yields a set of potential terms for the field. This method provides us with a uniform distribution of terms and ranks them according to their frequency of use.

Finally, the essential task is to compile the lexical knowledge base for the dictionary. The idea is to build a data base that can host the terms, a representative and diverse amount of their definitions and other relevant lexicographical data. In order to accomplish this, we apply to the corpus the ECODE tool, a system for the automatic extraction of useful information to describe the meaning of terms (Alarcón et al, 2008). This information is displayed by ECODE in structures called definitional contexts (DCs), i.e. text fragments where an author explicitly defines a term. The ECODE method, a pattern based approach for extracting definitional contexts in Spanish specialised texts, is based on the search of definitional verbal patterns related to four different kinds of definitions: analytic, functional, extensional and synonymic. The system includes three modules related to the extraction of definitional verbal pattern occurrences, the filtering of non-relevant contexts, and the identification of constitutive elements: term, definition and definitional verbal patterns. This system is used to extract the DCs from our corpus. We also use the WEB to obtain a huge amount of DCs used by diverse authors with different levels of expertise.

With the integration of all these tools and the method we propose, a valuable resource can be obtained that can be applied to the Mexican variant of Spanish in order to extract semiautomatically a wide range of terminologies. As we explained, if this method is used for term extraction and then implemented for automatic definition extraction, we can provide an automatic lexicography tool for the specialist in this area.

References

Introduction

The importance of word combinations is widely recognized in several domains, such as natural language processing (Smadja, 1993; Evert, 2005), lexicology (Ježek, 2005), lexicography (Benson, 1990; Cowie, 1981), language acquisition (Nesselhauf, 2005) and language teaching (Nation, 2001). Despite the lack of consensus on the definition of collocation, collocational competence is generally recognized as playing a key role in the linguistic competence of native speakers; while they have the ability to recognize such combinations as the most appropriate way of expressing a given concept, L2 learners who lack collocational competence often produce unacceptable combinations (Nesselhauf, 2005); this notably happens with collocations that are based on conventional linguistic behaviours rather than on semantic restrictions (fare una domanda, ‘to make a question’; Jaén, 2007).

Description of the project

This paper describes the Dici, a corpus-based Dictionary of Italian Collocations. The dictionary is a computational tool which aims to support the processes of training, testing and improving the collocational competence of students of Italian as a second language. The paper focuses on the integration of the Dici with an online learning environment specifically designed for Italian as a second language courses. The growing availability of natural language processing tools represents a great opportunity for online language learning environments, which can benefit from a vast quantity of structured linguistic data, typically extracted from corpora and stored in databases or dictionaries, and innovative computational methodologies (statistical and/or rule-based).

The Dici is therefore one of the computational resources integrated in April, our personal language learning environment (Ambiente Personalizzato di Rete per l’Insegnamento linguistico). It is a corpus-based dictionary (collocations are extracted from multi-genre, balanced corpora of Italian); it relies on statistical methodologies (collocations are ordered by mixing distinct criteria such as frequency, dispersion in the different textual genres represented in the corpus, and statistical measures; Krenn & Evert, 2001) as well as on a phraseological approach (Nesselhauf, 2005), 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.

The Dici contains a list of the most significant Italian collocations, ordered by frequency. It also provides the following structured information connected with lexical, syntactical, semantic, contextual and statistical aspects of collocations:

1. Grammatical category of the collocation.
2. Different syntactic configurations of collocations. Data must be grammatically analyzed to allow the recognition of collocations in a text.
The dictionary is therefore connected with a part-of-speech tagger, which 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. This capability of analyzing the syntactic structure has another advantage: learner errors can pertain to the syntactic configuration of the collocation rather than to its lexical features (for example *fare la confusione*, with the article, instead of the correct combination *fare confusione*, without the article).

3. Collocation type (following the classification proposed by Granger & Paquot (2008)).

4. Definition, extracted from existing dictionaries.

5. Context (samples of authentic text in which the collocation occurs, extracted from corpora).

6. Frequency in the entire corpus.

7. Frequency in different text typologies.

8. Dispersion in different text typologies (Hlaváčová & Rychlí, 1999).

**Integration with learning environment**

A great amount of research has been carried out in the last decades in the field of collocations, and more specifically of the creation of dictionaries of collocations (Santos Pereira & Mendes, 2002; Alonso Ramos, 2003, just to mention two recent examples for languages other than English). In this context, the *Dicí* project finds its specificity in its being aimed at the integration within an online learning environment. This means that in a specific area of our learning platform, devoted to the study of lexicon, 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 *Dicí*. Some of the features of the dictionary in its integration with the online platform are:

1. to automatically recognize and highlight collocations in written Italian texts (receptive level);

2. to provide a second language guided writing tool, in order to train collocational competence (productive level);

3. to generate collocation tests aimed at assessing the collocational competence of second language students (teacher tool).

The paper will present examples of this integration taken from a test version of the dictionary, that contains the verb + noun sequence of the collocations.

**References**


Towards a Multifunctional Electronic Dictionary Using a Metamodel of User Needs

Dennis Spohr
IMS, Universität Stuttgart
spohrds@ims.uni-stuttgart.de

The aim of this paper is to provide details on the definition of a model for a Multifunctional Electronic Dictionary (MED). The term “multifunctional” is borrowed from literature of the late 1980's and early 1990's on reusable lexical resources (see Heid & McNaught, 1991; Bläser et al., 1992) and has been discussed in great detail in Heid (1997). According to these sources, a multifunctional lexical resource is one that serves different types of users with different lexicographic needs as well as applications of natural language processing (NLP). It is thus in opposition to application-specific lexical resources, which have been designed to serve one particular purpose or a small set of predefined ones (e.g. subcategorisation lexicons in Lexical Functional Grammar), and which often require considerable effort when trying to adapt them to unforeseen functions or applications.

In more technical terms, multifunctionality refers to the ability to extract different dictionaries from a common lexical database, i.e. to provide different views on the lexical data. Each of these views comes with its own lexicographic needs, and thus the definition of an MED presupposes a clear understanding of the functions it is to serve, in addition to providing an efficient way for extending it to cover other functions. Since dictionaries should be conceived of with the needs of users in particular usage situations in mind (e.g. text reception vs. production; cf. Tarp, 2008), what is necessary for an MED is a formal tool that states which elements of the lexical description are needed in a particular situation and how they are presented to the user. Here, the central question is how such information is to be encoded, i.e. directly in the lexical database or on a separate layer on top of the dictionary. Here, we advocate a modular view on the dictionary: instead of storing the relevant information in the database, we devise a metamodel of user needs that provides the formal means to relate elements of the metalanguage and the object language with metalexicographic descriptions, thus allowing for the definition of specific user (and NLP application) profiles. This offers a powerful and extensible means to turn a shared source of lexical data into an electronic dictionary that is capable of accommodating multiple lexicographic functions.

Without going into too much technical detail, we will assume lexical data that have been structured and implemented in an XML-compatible format, such as the Semantic Web formalisms RDF(S) (Resource Description Framework Schema; Brickley & Guha, 2004) and OWL (Web Ontology Language; Bechhofer et al., 2004), or any other custom XML format. As was discussed in Spohr & Heid (2006) in the context of OWL-based bilingual dictionaries, such formalisms can make use of XML namespaces and uniform resource identifiers (URIs), which enable lexicographers to organise and reference the different components of a lexical resource in a modular way (cf. Spohr, 2008).

Transferring the core ideas of this approach to an MED, we arrive at an architecture that provides a separate layer for user and NLP needs. This layer is
situated on top of the model that describes the dictionary data and their structure, such as the different types of lexical entities, their relations and properties, and the necessary data categories.

![Diagram showing user profiles in relation to dictionary model and data](image)

**Figure 1:** User profiles in relation to dictionary model and data

As illustrated in Figure 1, the *user profile* layer contains different profiles that differ wrt. their lexicographic function. For example, one profile specifies the needs of users whose lexicographic usage situation is text reception, while a different profile describes the needs of users who want to produce a text in their mother tongue. As was mentioned above, Tarp provides a theory of such lexicographic functions, along with the primary and secondary needs associated with each function (cf. Tarp, 2008: chapters 4-6). Our user profiles directly implement these needs on the basis of Tarp's account by specifying, for each type of information in the dictionary, whether it is a primary or secondary (or possibly further) need of a user in that particular situation. While the question as to how the status as primary or secondary need is reflected in the dictionary entry is a topic of its own, it is important to note that this modelling of user needs enables uniform treatment of each kind of need across different profiles.

In addition to the status of a piece of information, each profile specifies multilingual *labels*. On the one hand, this enables users to navigate between different presentation languages within a given profile. On the other hand, it allows lexicographers to define different labels for the same entity in different profiles. For example, while entities of the type *Collocation* may be presented to lexicographically proficient users as “Collocations” in English or “Kollokationen” in German, it might be more appropriate to present them as “Typical word combinations” or “Typische Wortverbindungen” to a non-expert language learner. In the proposed setting, the
status and labels which apply in a given profile can be addressed by assigning to each profile a different XML namespace prefix, e.g. recmt: and prodmt: for text reception and production in the mother tongue respectively. The relevant kinds of information can then be easily retrieved on the basis of the selected user profile, so that users are able to dynamically switch between different presentation languages and profiles, with the appropriate dictionary entry being generated on the fly.

Thus, the approach outlined in this paper contributes to current lexicographic research in several respects. On a representational level, it provides details on a highly modular architecture of an MED, implementing a clear separation of object language(s), metalanguage and presentation language(s). Moreover, it demonstrates the use of up-to-date formalisms and technologies developed within the Semantic Web, whose suitability for computational lexicography has recently been attested in Görz (in press). In particular, we show how these formal means enable us to define the model for a MED for human usage and NLP scenarios. Finally, with the strong emphasis put on the modelling of user needs as a separate module of its own, a flexibility of presentation is achieved that shifts the notion of dictionary entry from a static entity to a dynamic one that is generated automatically at runtime, based on the specific needs of the users in their current usage situation.

References


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

Based on the concept of Usuelle Wortverbindungen (UWV, Steyer, 2000; 2004; 2008), we regard multi-word expressions as conventionalized patterns of language use that manifest themselves in recurrent syntagmatic structures. This concept encloses not only idioms and idiosyncratic structures, but all multi-word units which have acquired a distinct function in communication. Our approach focuses strongly on real-life usage, pragmatics and context. It also takes into account that MWE are very flexible and variable and that there are numerous interrelations between them.

In detecting as well as describing these units we work bottom-up in a strongly corpus-driven way (Sinclair, 1991; Tognini-Bonelli, 2001; Hanks, 2008; Brunner & Steyer, 2007), taking iterative steps to arrive at conclusions about language use. At the moment, our focus is on developing ways to handle corpus and collocation data for lexicographical presentation targeted at learners of German or linguists interested in patterns of usage. The resulting structure is inspired by statistical data, but human generated.

The basis of our research are collocation profiles, computed with a sophisticated method (Kookkurrenanalyse, Belica, 1995; KLb2008) from DeReKo (Deutsches Referenzkorpus, KLa2008), the largest collection of written German available today which has approximately three billion word tokens and is located at the Institute for the German Language (IDS). In this experimental stage, we use DeReKo as it is.

Starting from a target word form, we manually explore the statistical collocations, using a method of surface pattern matching, which allows the grouping of instances of usage in the form of Key Word in Context (KWIC) lines. In the next step, these instances are bundled into groups which represent specific types of lexical realisations of an MWE. Corpus research shows clearly that the surface form of an MWE is nearly always subject to variation and the lexical realisations (LR) are a way of illustrating this by grouping the relevant corpus data. 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 realisations can exist, called core variant LR in our model. In addition, we define extension LR, 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 are context LR, which highlight word forms which typically appear close to the MWE realisation without being part of its structure.

Each MWE in our model is described by a bundle of lexical realisations, guiding the attention of the recipient to collections of real corpus data which illustrates a certain aspect of the MWE’s structure and/or usage. The following
example illustrates the LR structure for the MWE “Musik in den Ohren”.

(1) **Core LR:** Musik in den Ohren [music in the ears]

**Core variant LR:** Musik in X (possessive pronoun/genitive phrase) Ohren [music in X’s ears]

**Extension LR:** Musik in X Ohren sein [to be music in X's ears]

**Extension LR:** wie Musik in X Ohren sein [to be like music in X's ears]

**Extension LR:** wie Musik in X Ohren [like music in X’s ears]

We also explore gaps in the recurrent structures and provide two types of filler lists: firstly, automatically generated lists from the corpus data for every gap in our patterns, secondly, manually generated lists of typical fillers for gaps which show regularities. These lists give important insight into the paradigmatic variability of MWEs.

The MWEs themselves are arranged in groups, resulting in more abstract forms with variant components, which share a common communicative function. Frequent combinations of MWEs as well as related communicative functions of different MWEs are pointed out. An extract of an MWE group hierarchy is presented here:

(2) aus+ADJEKTIVE+Gründen [for ADJEKTIVE reasons]

aus+ADJEKTIVE\_DOMAIN+Gründen [for ADJEKTIVE\_DOMAIN reasons]

MWE: aus politischen Gründen [for political reasons]

MWE: aus finanziellen Gründen [for financial reasons]

MWE: aus technischen Gründen [for technical reasons]

aus+ADJEKTIVE\_QUALIFICATION\_POSITIV+Gründen [for ADJEKTIVE\_QUALIFICATION\_POSITIV reasons]

MWE: aus guten Gründen [for good reasons]

MWE: aus nachvollziehbaren Gründen [for understandable reasons]

MWE: aus triftigen Gründen [for cogent reasons]

aus+ADJEKTIVE\_QUALIFICATION\_NEUTRAL/NEGATIVE+Gründen [aus+ADJEKTIVE\_QUALIFICATION\_NEUTRAL/NEGATIVE+Gründen]

MWE: aus unbekannten Gründen [for unknown reasons]

MWE: aus unerklärlichen Gründen [for unexplained reasons]

MWE: aus unerfindlichen Gründen [for unclear reasons]

In this example, a communicative function which is shared by all its child MWEs can be attributed to the abstract MWE aus+ ADJEKTIVE\_DOMAIN+Gründen: Using this pattern makes the actions that are explained seem official and at the same time allows the speaker to be vague about the reasons for these actions by using the less specific plural form (“Gründen”/reasons) which is mandatory for its structure.

The results of our research are represented in fields of MWEs, each centred on
a specific word form. The hierarchical structures and interrelations between the different units are realized in a hypertext format and presentation gives direct access to structured corpus data. All levels of description are enriched by lexicographic comments like the description of meaning and usage in the corpus. Thus two forms of reception are possible: Firstly the structure allows the reconstruction of the typical usage of MWE from the corpus data and provides a complete documentation of our interpretative method. Secondly the narrative comments allow a reception similar to that of traditional lexicographical products. The first version of Fields of MWE, one centred on forms of the word Grund and one around forms of the word Ohr are available on the internet, accessible from our site “Wortverbindungen online” (Steyer & Brunner, 2008a; b).

Though developed in an experimental research context, we believe that our approach can give valuable impulses to lexicographic practice: Our model presents a new approach by including corpus data not only as illustration, but as the basis of description, and offers structured access to real-life data, taking advantage of the options of electronic hypertext format.

References


Beyond Lexicocentrism: The Challenge of Complex Architectures in eLexicography

Toma Tasovac
Center for Digital Humanities (Belgrade, Serbia)
ttasovac@transpoetika.org

The promise of eLexicography stems not only from the transformation of the production medium, but also from the technological feasibility of representing linguistic complexity. Even though modern lexicography is unimaginable without computer technology (Knowles, 1989; Meijs, 1992; Hockey, 2000), the sheer use of computers in producing a dictionary or delivering it electronically does not automatically transform a dictionary from "a simple artifact" to a "more complex lexical architecture," to use Sinclair's (2000) formulations.

Calling dictionaries “simple artifacts” is itself a rhetorical oversimplification: there is nothing simple about a dictionary -- whether we look at it as a material object, cultural product or a model of language, yet the overall structure of dictionaries as extended word lists has not changed in centuries. Admittedly, a great deal of factual information is packed into a prototypical lexicographic entry, but a defined term often remains in isolation and insufficiently connected or embedded into the language system as a whole. This is what Miller refers to as the “woeful incompleteness” (Miller at al.) of a traditional dictionary entry, and what Shvedova sees as its “paradoxical nature” -- dictionary entries tend to be “lexicocentric” while language itself is “class-centric” (Шведова, 1988).

Transpoetika is a collaborative, class-centric, bilingualized Serbian-English learner's dictionary based on the machine-readable semantic network of the Princeton Wordnet (Vossen, 1998; Fellbaum, 1998; Stamou et al., 2002; Tufis, Cristea & Stamou, 2004). The project is currently under development at the Center for Digital Humanities (Belgrade, Serbia) and is part of a scalable, web-based, digital framework for editing and publishing annotated, fully-glossed study editions of literary works in the Serbian language, primarily aimed and students of Serbian as a second or inherited language.

By treating all of its entries as members of semantic classes, and by placing these classes in hierarchies, Transpoetika maps linguistic meaning into a dynamic, logical metasystem. This, in turn, makes it possible to see a dictionary entry not only as a depository of lexicocentric information about a given headword, but as a point of departure for the user's exploration of the semantic relations between words and their senses (for instance synonymy, antonymy, hypernymy, hyponymy, entailment etc.) By explicitly encoding lexical and semantic relations, our Wordnet-based dictionary makes it possible -- and easy -- for language learners to access words they do not necessarily know by making it trivial to execute database queries of the type “What are parts of *?” or “What does * entail?” etc.

At the same time, our dictionary extends the scope of the two mutually aligned Wordnets by providing methods of encoding a host of important lexicographic properties for the Serbian entries: detailed grammatical information, full accented declensions and conjugations, usage labels (dialectological, temporal, functional), and, when appropriate, stylistic distinction among members of a single synset. This
way, Transpoetika will be able to meet the requirements of a full-fledged L2-L2-L1 dictionary (Wiegand, 1985) without abandoning the architectural rigor of Wordnet’s semantic network.

The entire project is firmly rooted in our belief that the management of complex semantic relations and the widening of distribution channels are some of the most important challenges facing electronic lexicography today. Electronic dictionaries should be more than digitalized versions of lexicocentric print texts. Instead, they should be built upon more ambitious lexical and semantic foundations that reflect not only individual lexical items, but their complex mutual interactions as well.

The future of electronic dictionaries lies in their detachability from physical media (CD, DVD, desktop applications) and static locations (web portals). Transpoetika has been designed so that it can be deployed as a web service and therefore linked from and applied to a variety of textual sources. Portions of the project, such as the Serbian Morpho-Syntactic Database (SMS) already function as a web service internally and will be made public and free once the sufficient funding for the project has been secured. The same applies to the dictionary on the whole, which could become an important cornerstone in the development of the Serbian semantic web.

With foreign-language teaching increasingly focusing on the role of the lexicon and a need to move the pedagogical focus from lexicalized grammar to grammaticalized lexis (Nation, 1990; Lewis, 1993; Carter, 1998; Singleton, 1999; Lewis & Conzett, 2000; Schmitt, 2000; Nation, 2001), Transpoetika could also become an important pedagogical resource for a lesser taught and lesser resourced language such as Serbian.

References


Fine-Tuning a Common-Sense Classification in the ANW Dictionary

Carole Tiberius & Fons Moerdijk
Institute for Dutch Lexicology (INL)
The Netherlands
{carole.tiberius,fons.moerdijk@inl.nl}

Introduction
In this paper we discuss the semantic classification which is used in the ANW dictionary project. ANW stands for Algemeen Nederlands Woordenboek (General Dutch Dictionary), a comprehensive online scholarly dictionary of contemporary standard Dutch in the Netherlands and in Flanders, the Dutch speaking part of Belgium. The project runs from 2001 till 2019 and the first results will be published on the web in 2009. Ultimately, the dictionary will contain 80,000 headwords with a complete description and about 250,000 smaller entries. There will not be a printed version of the dictionary.

The ANW is a very informative dictionary with an abstract entry structure consisting of hundreds of elements and subelements. One of its main innovations is a twofold meaning description: definitions are accompanied by ‘semagrams’ (Moerdijk, 2008), the representation of knowledge associated with a word in a frame of ‘slots’ and ‘fillers’. ‘Slots’ are conceptual structure elements which characterise the properties and relations of the semantic class of a word meaning. On the basis of these slots specific data is stored (‘fillers’) for the word in question. The abstract structure schema is called a ‘type template’, whereas semagram refers to a ‘type template’ populated with concrete word data. Each semantic class has its own predefined type template with its own slots.

Motivation
The ANW adopted a bottom-up approach for defining the semantic classes and corresponding type templates. There are numerous ontologies and semantic classifications available, but these classifications are often too elaborate or they contain too many or too few semantic relations for the particular purpose of the ANW dictionary (e.g. SUMO¹, Cyc²). The ANW wanted a classification geared towards lexicographic description and based as far as possible on linguistic foundations rather than a classification based on a division of words over various social domains. In addition, the classification had to be relatively transparent such that it could also be used in the dictionary’s search part supporting onomasiological queries. In the ANW, the user can perform onomasiological queries by typing in a definition, summing up words that spring to mind or by using a guided search based on the semagram. In this case, the user is asked to select a category (the semantic class or subclass) from a menu (is it a thing, a person, an animal, a vehicle, etc.?) after which the feature slots of the type template for that category appear on the screen and the user is asked to fill in value(s). It is obvious that for this to be successful a limited amount of classes and features is needed in order not to confuse the user.

¹ http://www.ontologyportal.org
² http://www.cyc.com
Thus, the ANW system is very much in line with what Cruse (2000) calls a common-sense hierarchy. However, as it should also be possible to use the classification in computational applications, an effort was made to keep the hierarchy clean.

Method

The current classification of nouns, verbs and adjectives results from manual analyses of 7000 definitions taken from two different dictionaries which were subjected to various critical clustering and refining tasks starting from the ‘genus proximum’ in the definitions. For nouns this process resulted in 266 semantic classes, subdivided into 49 main classes. For verbs, 195 classes were obtained in 22 main classes and for adjectives 9 main classes were distinguished.

After identification of the classes and subclasses, type templates with class-specific features were compiled for each by adding together all the analysed features for all the words belonging to that particular class.

The resulting type templates formed the basis for the completion of the first 1000 semagrams. These semagrams together with the system resulting from the manual analyses were then subjected to some clustering analyses and compared to related work on ontologies and semantic classifications, in particular WordNet (Fellbaum, 1998), CISLEX (Langer, 1996), DanNet (Pedersen et al., 2006) and SUMO, and relevant work in psychology such as McRae et al. (2005) and Garrard et al. (2001), resulting in a system with the following characteristics:

- The semantic hierarchy has a (relatively small) number of generic concepts (19 for nouns), and each one is treated as the unique beginner of a separate hierarchy. Thus the resulting structure is not a complete tree, but a set of disjoint hierarchies.
- A clear distinction is made between type relations and other semantic relations such as roles and parts, which are dealt with in terms of features in the type template of a semantic class.
- Each level in the semantic hierarchy contains two slots separated by a # character. Before the hash character we aim for a ‘fixed’ type hierarchy, whereas data resulting from hierarchical classifications based on one specific feature of a class will be put after the # character. This way we get a kind of double-layered hierarchy. For instance, when classifying trees, the distinction whether a tree has needles or leaves, i.e. whether it is a conifer (‘naaldboom’) or a deciduous tree (‘loofboom’) will occur behind the hash character as it is a classification based on one specific features of trees, namely the shape of the leaves. The classification of plants/vegetation in trees, flowers, bushes etc., however, will occur before the hash character.
- The type templates provide a very fine-grained semantic feature description, more so than in most existing systems. For use in onomasiological queries, the most distinctive features per class are identified and given priority in the lexicographic description.

1 In total, there are 166 features divided into 22 main groups.
The lexicographers give the characterisation of the feature slots in terms of short statements about the headword. These sentences are unified using a clustering analysis of already completed semagrams and can, as such, be used to express the different semantic relations.

The result is a transparent, easy to use, principled system for the representation of knowledge associated with a word in a systematic way.

References

Building Multilingual Legal Lexicons

Daniela Tiscornia
Italian National Research Council- Institute for Legal Theory and Information Techniques (ITTIG-CNR)
tiscornia@ittig.cnr.it

When examining the legal vocabulary, we encounter two different types of semantic information associated with elements from legal text. On the one hand, there is ontological structuring in the form of a conceptual model of the legal domain, consisting of a complex structure of concepts, forms and abstraction from legal textual material. On the other hand, there is a vocabulary of lexical items that lexicalize concepts, which are not necessarily restricted to the legal domain, and are associated with specific linguistic information. We can therefore argue that 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, that guide its evolution and guarantee its consistency. Both are conditioned by the social dimension in which it is placed, whereby they dynamically define and fix their object in relation to a continually evolving social context.

To make the scenario even more complex, terminology used in the various legal systems, both European and national, expresses not only the legal concepts which operate there, but further reflects the profound differences which exist between the various systems and the differing legal outlook of the 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.

This complexity poses crucial problems for accessing and handling legal information, as traditional search engines for legal information retrieval do not include legal knowledge into their search strategies, on which to perform conceptual query expansion and cross-lingual retrieval.

The building up of computational tools aimed at bridging the gap between terminologies and concepts requires an ontology-based methodology, since the representation of legal concepts in ontological frameworks makes explicit assumptions about legal knowledge, helps the understanding and sharing of legal concepts, mediates storage of legal rules and supports reasoning on them. In particular, methodologies based on frame semantic and ontology learning techniques seem to be the most promising way to fill the gap between dogmatic conceptual models and the lexical layer extracted from texts. Such a model expresses 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.

Several issues linked to the definition of such a methodology will be outlined in the oral presentation both from the theoretical perspective, underlying the conceptual model, and from the practical realization, which relies on recent and
ongoing projects.

From the first point of view several issues will be discussed, among which:

- layers distinctions in legal discourse
- the multiple perspectives of legal texts analysis
- the notion of legal concept
- overlapping between domain and semantic properties
- the role of contexts and of definitions
- the top down vs bottom-up (corpus-based) approach.

From the perspective of practical implementation, experiences and lessons learned from two European projects, LOIS (Lexical Ontologies for legal Information Sharing, EDC 22161, 2004-2006) and Dalos (Drafting legislation with ontology-based support, eParticipation, 2006-2007) will be presented, concerning:

- the overall design and the formal architecture of a semantic repository
- concepts selection and clustering
- multilingual matching
- properties derivation
- ontology-based frames detection

Finally, still unresolved aspects will be discussed and future research directions will be presented.

References


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, Germany
sabine.tittel@ urz.uni-heidelberg.de

New lexicographical projects start with certain advantages: in their case, computer scientists don't need to deal with the problem of how to dress a weighty tome (a long-established dictionary) in fancy new clothes. Far from it: 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'ancien français (DEAF), the Etymological Dictionary of Old French, compiled in Germany at Heidelberg University and under the aegis of the Heidelberg Academy of Sciences.

We need to visualize the substantial corpus of our 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 before the new process began.

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 and to the key challenge of twenty-first century lexicography. In other words: the transformation of lexicography into eLexicography, a process which implied a new dictionary architecture, and the successful bringing-together of both new elements, and of those which have been scientifically established for decades.

Thus the challenge on which 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 (we obviously enlisted professional computer science assistance, see below) but rather in the interface between lexicography and computer science.

The transformation necessitated:

(1) an entirely new interior;
(2) and an entirely new exterior.

Some facts...

... about the source material and corpus of the dictionary:
• vast quantity of Old French sources: the dictionary's bibliography (see below)
• an “open” corpus: all available editions of all texts, manuscripts,
secondary literature, dictionaries, etc.
- 1.5m handwritten slips (fiches) with heterogeneous information leading to 12m attestations and serving as an entrée into the sources (the "ad fontes" principle: always go back to the sources)

... about the dictionary:
- ongoing composition for over 40 years (volumes G to K [1974 - 2008]: 10,217 entries, 4,099 columns, 533 pages of printed indices)
- the dictionary's critical and annotated bibliography DEAFBibl (the most comprehensive of its kind, the industry standard work used also by other dictionaries, journals, etc. [ANDEI1, DMF 20092, CCFM3, Nouveau Corpus d'Amsterdam4], references to more than 6,000 sources), 2007, 1,031 columns5
- historical dictionary: Old French, taking into account its further development to Modern French and all vernacular languages of medieval Europe including Mediaeval Latin
- etymological dictionary: involving Latin, Greek, Germanic languages, Hebrew, Arabic, etc.
- complex and dense dictionary entry structure due to scientific content: information is presented in a maximally space-saving manner whilst preserving a maximal quality

Excursus: "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 articles’ 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.6

The DEAF is characterised by a very complex article structure. The article is composed of a main entry with subentries which each consist of: 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); graphical variants (with Old French scriptae, attestations which may differ from those of the semantic article section, text and manuscript dating, inflection related to Old French two-case system); the semantic section (divided into numerous

---

2 Cf. Analyse et traitement informatique de la langue française ATILF - Nancy Université & Centre National de la Recherche Scientifique CNRS (2009).
3 "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.
4 Cf. Stein, A. et. al. (2006).
6 A point of discussion ignored in this paper is the new articles' 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.
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 potentially be found 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, the first three textual references to each of the first ten source texts in chronological order are referred to. More textual references to one source text are condensed as "etc.", more source texts as "etc.etc.". Relevant references to additional attestations and the relevant information which they contain 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, of parts of speech, etc.). This principle of condensed information also works in part without explicit abbreviations/indicators, relying purely on the user's knowledge of the article structure.

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 combination of (and the sheer number of) the DEAF's characteristics.

(1) The new interior is represented by a new and complex editorial system which allows for a technically supported and time-saving production of dictionary entries, via a completely new system which has been designed for the use of the editorial team.

Some facts...

... about the dictionary editorial system and the technical solution (in cooperation with the “Institute for Program Structures and Data Organization” IPD, Prof. Dr. Dr. h.c. Peter C. Lockemann, University of Karlsruhe, Germany):

- MySQL-Database, user interface WicketFrameWork, Hibernate, Databinder
- digitized data (XML) in the form of 1.5m slips and 6,000 bibliographical references
- lemmatization of slip data (Java; complicated by spelling variation in Old French; based on 120 phonetic rules regarding diatopic variation in Old French and its historical development from Latin)
- 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

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 freedom which humanities scholarship always requires. 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). To achieve this we agreed to a compromise
which resulted in the combination of the benefits of two ostensibly contradictory elements: i) benefits of electronic work-flow support with fixed structures and automated data management, together with ii) the benefits of free-text editorial input, with the facility of semantic markup.

But the tool of free-text input has implicit disadvantages (mainly due to monetary restrictions: we could have done better with more time and money...), and our most substantial acknowledgement of the freedom of the humanities was that we abandoned an automated convergence of the structured data managed by the electronic work-flow support, and the data presented in the published articles. This decision also inevitably entails a higher risk of errors which are not detectable by automatic data synchronisation.

Implicit, too, in preserving the freedom of the humanities was the question of whether it is (technically) possible to reproduce one-to-one 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. Positive, that is, if computer science follows the route set out for it and 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.

(2) The new exterior shows a dictionary offering on-line searchability and allowing for flexible research queries, a completely new product for the dictionary user.

Some facts...
... about the on-line version (starting 2009/2010)
• 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
• access to dictionary information separate from the dictionary's structure: a versatile system of 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\(^1\), the Woordenboek\(^2\) or the DMF 2009
• publication of the entirety of the DEAF's raw material (slips, etc.) offers research possibilities beyond the dictionary's own contribution\(^3\)
• the integrated on-line publication of the DEAF bibliography, including search functions

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.

---

\(^1\) Cf. Murray, J. A. H. (1888-1928).
\(^2\) Cf. de Vries, M. & Te Winkel, L. A. et al. (1882 →).
\(^3\) This raw material is not to be confused with digital text sources consultable for example via the ANDEI.
Consequently, they cannot be relocated elsewhere.

Examples...

1. ... from the semantic section, which consists of main- and sub-senses: There are sub-senses, which feature a complete definition, and which are thus self-sufficient. But numerous sub-senses are defined via "idem" and additional information (e.g. "id." used as a metaphor) and thus cross-refer to the definition of a previous sense (main or sub-sense). (A similar concept is used by the MED1, LEI2, TLF3 and the DMF 2009; self-sufficient (sub-) senses are defined by the AND4, Gdf5, OED, Goethe-Wörterbuch6, Woordenboek, etc.)

2. ... from the graphical variants' section: without being expressed explicitly, the indications of Old French scriptae of the graphical variants apply until such time as a different scripta is indicated or until the scripta indication is explicitly cancelled.7

3. ... from 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.

The basis of this space-saving concept is the principle of condensed information which partly works with conventional abbreviations/indicators like "idem", "etc.", "etc.etc." (see above), and partly without the use of such indicators, but relying on the user's knowledge of the structure of articles. This principle implies the disadvantage that the intuitive use of the dictionary only covers the information given in an explicit way but not the information included 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, in other dictionaries, access (via search functions) to this implicit information (if at all contained within a static article structure) does not always work smoothly, 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, of course, to the entry veine and being the mediaeval term for what we now define as the aorta. Not knowing that, and searching 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 alone.

The DEAF has set as a target that the condensed external form of the dictionary as conceived as a static dictionary will not affect the searchability of the information which is explicitly and implicitly given in the dictionary's articles.

This brings us to the final question: is eLexicography determined by technical potential, that is by "e", or by lexicography?

---

7 With "s.l." meaning "sans localisation" (sine loco).
Both 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.

In the case of the DEAF the transformation process was complicated by the range of possibilities of supplying, arranging, summarizing, or (where this follows the article's scientific concept) of omitting information. The dictionary, we would respectfully suggest, exceeds most others in its multiplicity. And it was philosophically imperative for the lexicographers concerned that we should capture the idea of the freedom of this range of options, rather than adopting completely fixed structures and wholly automated data management.

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 (to the IPD, see above) saved the project. The cooperation is now far more balanced.

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


Access to Multiple Lexical Resources at a Stroke: Integrating Dictionary, Corpus and Wordnet Data

Lars Trap-Jensen, Henrik Lorentzen & Jørg Asmussen
Society for Danish Language and Literature
ltj@dsl.dk, hl@dsl.dk, ja@dsl.dk

The non-linear organisation of hypertextual elements has opened new possibilities and ways of seeking information. Among other things, users can query different types of data simultaneously, resulting in easier and quicker access to information. One lexicographical perspective of this is that re-arrangement of existing data can lead to the creation of new and useful data. On the other hand, as the possibilities grow and data become increasingly complex, producers and developers of lexicographical tools are faced with a growing challenge of balancing the new possibilities with a manageable and appealing design.

In this paper we present 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.

While it is true that these elements are of a somewhat diversified nature, they have more in common than one might first think. The 28-volume Dictionary of the Danish Language was completed in 1955, with five supplementary volumes appearing 1992–2005, and The Danish Dictionary was published in six volumes 2003–2005. Despite their historical differences, they share a number of characteristics, having been compiled by the same institution, The Society for Danish Language and Literature, and within the same descriptive lexicographical tradition. In terms of data, the remaining resources are even more closely connected, The Danish Dictionary being the first – and so far the only – dictionary of Danish to be compiled as a corpus-based dictionary, and the corpus component of www.ordnet.dk comprising all the non-restricted, written texts on which the dictionary was based, as well as more recent texts added later. After its completion in 2005, data from The Danish Dictionary were used to construct a Danish wordnet following the well-known models of Princeton WordNet and EuroWordNet. So although the components of the site appear different at first glance, the underlying data are sufficiently uniform to allow cross-component exploitation in a way that is, we think, both innovative and relevant by fulfilling important user needs.

In our paper we will focus more specifically on three aspects: First, how are data exploited across the components? This section deals with onomasiological queries in the dictionary based on wordnet data and compares the approach with similar thesaurus functions in other dictionaries, e.g. the SMART Thesaurus of the Cambridge ALD (on CD-ROM) and the 2009 version of Macmillan English Dictionary. To our knowledge, dynamic extraction of corpus information has not been utilized very much in the field. An early attempt was the “Example Bank” of LDOCE, 4th edition, but the project most similar to ours is probably the German DWDS which integrates dictionary, corpus and thesaurus information at one site, though at present based on less uniform data. In this connection, we explore the possibility of supplying dictionary entries with additional examples as well as frequency information.
Secondly, we account for the adjustment of data structure that is required by the desired search possibilities. Particular attention will be devoted to inflections and variant forms, cross-references and multi-word units. Finally, questions of design and functionality are addressed, and the most important choices are discussed and compared with similar projects such as DWDS in Germany, OED, LDOCE and others in the UK.

References


NLP Tools for Lexicographic Applications in Modern Greek

Christos Tsalidis, Elena Mantzari & Mavina Pantazara  
Neurolingo L.P.  
{tsalidis, emantz, mavina}@neurolingo.gr

The aim of the presentation is to demonstrate the infrastructural software tools developed by Neurolingo L.P. in order to support lexicographic NLP applications and related computer-assisted activities. Neurolingo’s full list of products comprises a number of electronic language resources for Greek (i.e. dictionaries, thesauri, ontologies), language tools (i.e. lemmatizer, authoring/proofing tools, lexicographic databases and editors, corpus management systems), and software systems for NLP applications, such as multi-word term recognition, text mining and information extraction.

The demonstration will focus on the following software systems used for the development of lexical resources:

- **The “LexEdit”** is a language independent software used for the compilation of computational morphological lexica. The system supports the encoding of morphosyntactic attributes (i.e. grammatical category, number, gender, voice, person, etc.), hyphenation, inflection, and morphematic compounding (i.e. stem, prefix, affix, infix, suffix) of the lexemes.

- **The “Lexikografos”** is an integrated environment for editing and authoring of monolingual and bilingual dictionaries and thesauri of synonyms and antonyms. The system can manage XML files through a user-friendly interface, which further allows cross-referencing integrity between entries of the dictionary, spell-checking of the data, automatic up-dating of sense numbers, entry previewing (full or partial), grouping of lemmas using filtering, generation and display of hyperlinked related lemmas, searching on the headwords list and/or other fields of an entry, selection of the information to be exported. Moreover, the system is designed to support the production of a printed PDF version of the dictionary keeping the formatting annotations through XSLT.

- **The “Kanon”** is a feature-based grammar formalism, which is used for the recognition of specific morphosyntactic patterns in the input text. 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 (NER) in the framework of text mining and information extraction, and grammar checking.

The above systems have been used for the development of Greek lexical resources, incorporating, thus, a number of features (e.g. hyphenation rules, inflectional paradigms, spell-checking functionality, syntactic patterns) to support the description of this language. Notwithstanding, all of the systems, with minor adjustments, can be reused for encoding data from other languages too. To give a quick overview of the existing Greek language resources: the morphological lexicon comprises ~90,000 lemmas, i.e. 1,200,000 word forms, the electronic dictionary of geographic names and toponyms of Greece ~10,000 lemmas, the electronic dictionary of biomedical terms ~10,000 lemmas, and the thesaurus of synonyms and antonyms ~22,000 lemmas.
Finally, we will present some application tools based on the above mentioned resources, such as the proofing/language tools (spelling checker, hyphenator, thesaurus/synonymms), which assist text production and text editing in Greek and are available for various word processors and desktop publishing systems, and the “Lexiscope”, a web-based language application tool which provides grammatical and semantic information about a Greek word or phrase, combining Neurolingo's hyphenator, spelling-checker, lemmatizer, morphological lexicon and thesaurus (http://www.neurolingo.gr/en/online_tools/lexiscope.htm).

References


Tsalidis, C., Orphanos, G. & Iordanidou, A. (2002). Οι ελληνόγλωσσοι υπολογιστές έμαθαν να συλλαβίζουν (;) [Did the Greek computers learn how to hyphenate (;)]. In Studies in Greek Linguistics, Proceedings of the 23rd Annual Meeting of the Department of Linguistics (pp. 901-911), Aristotle University of Thessaloniki, Thessaloniki, Greece.
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, tend 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.

One of the more 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’ view of polysemy in L2 words will in part depend on their L1, which might have similar polysemy for individual words. Divergent polysemy (see examples below) is likely to produce a particularly high learning burden.

1. L1 German Schatten > L2 English shade / shadow
2. L1 English river > L2 French fleuve / rivière
3. 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. (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.

To make such sophisticated vocabulary trainers possible, we need lexical databases that contain more information than the ones we have today. If we choose a bilingual lexical database as a starting point, information about word frequency and a
didactically sensible grouping and prioritizing of translational equivalents is necessary. For each repetition cycle, a small number of suitable example sentences illustrating the appropriate subsense should be available. 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 plus corpus examples might seem to be an dauntingly huge task, we should remember that we are dealing with learner language here. A few thousand lexemes are likely to be more than enough for the vast majority of foreign language learners.

The poster aims to describe a template for a lexical database needed for the type of intelligent vocabulary trainer sketched above. Such a database could be seen to have some parallels with learner’s dictionaries or bilingualized dictionaries, but while it should not be seen as 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.

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 insufficient.

Deciding on the appropriate subsenses for polysemous words will be one of the first difficult decisions to take 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. The problem is the divergence among these sources. To use an example from Atkins & Rundell (2008), 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 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 & Nation, 1993). All subsenses and their syntactic contexts will then need to be illustrated with a series of examples, graded by difficulty.

References


---

1 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.
Showing Phraseology in Context: An Onomasiological Access to Lexico-Grammatical Patterns in Corpora of French Scientific Writings

Agnès Tutin
LIDILEM – Université Grenoble 3 – Stendhal
agnes.tutin@u-grenoble3.fr

As outlined by Gledhill (2000) and Pecman (2004), phraseology is particularly prevalent in scientific writings and a good command of these prefabricated patterns is essential to write a high quality text. A large number of dictionaries and course books, especially in English, provide useful lists of this phraseology (see for example De Cock et al. (2007) in the Macmillan Dictionary), generally organized under semantic or rhetorical labels. Resources of this kind are nevertheless rarer in French. However, what non native speakers often need most is a large set of authentic corpus-generated examples, which illustrate the phraseology in its natural environment. In French, one such promising lexical database, the dictionary of Antidote (see Charest et al., 2007), links in a systematic way collocations and examples of these collocations in authentic texts. However, no semantic treatment is provided in this lexical database.

In this paper, we would like to argue for an onomasiological approach to phraseology, which accesses corpora via rhetorical and semantic functions in scientific writings. Our first linguistic treatments deal with lexico-grammatical markers of authorial position in scientific writings. This linguistic topic deals with how the author is represented through his or her production, how he/she refers to earlier work, situates the novelty of the findings within the field and evaluates her/his own work or peer work.

Our theoretical framework includes three linguistic levels:

- The rhetorical level deals with the different kinds of rhetorical functions, e.g. evaluating a scientific object, situating one’s own work in the research field …
- The semantic and conceptual level enumerates in a frame-like approach (e.g. Fillmore et al., 2003) the frame elements involved in a semantic frame. For example, the frame of ‘evaluation’ involves an agent, an object and a polarity (for polar evaluations). The frame of ‘scientific inheritance’ involves a scientific agent who “inherits” a scientific object from a scientific source, who stems from a school of thought, ideas, etc.
- The lexico-grammatical (and enunciatice) level indicates how the participants and the frames are linguistically expressed, for example, for axiological evaluation, specific adjectives like interesting or remarkable, while scientific objects evaluated are words like methods, approaches or results. For the ‘scientific inheritance’ frame, the scientific agent is generally the first person pronoun while the scientific source is a citation, and the scientific object belongs to a closed set of scientific artifacts (notion, concept, model …) (see examples 1 and 2).

1) **We** (scientific agent) **borrowed** (scientific inheritance) **the notion** (scientific object) of test and training sets from the theory of neural

(2) Following (scientific inheritance) the work (scientific object) of Edwards and Palsson (2000a, b, c) (scientific source), we (scientific agent) studied the effects of gene ...

Our model is implemented in a query system using a large corpus of varied scientific writings in the humanities, social sciences and “hard” sciences. The onomasiological access is organized in a hierarchical way, e.g. ‘evaluation’ is subcategorized into ‘axiology’, ‘importance’, ‘novelty’, ‘temporal’, ‘degree’ The lexico-grammatical patterns are integrated in grammars using regular expressions and syntactic dependencies parsed with the Syntex system (Bourigault, 2007) and will be implemented in an online user-friendly interface.

References

Tuebingen: Gunter Narr Verlag.
SciE-Lex: An Electronic Lexical Database for the Spanish Medical Community

Isabel Verdaguer, Natalia Judith Laso, Eva Giménez, Danica Salazar & Elisabet Comelles
University of Barcelona
i.verdaguer@ub.edu; njlaso@ub.edu; egimenez@ub.edu; dsalazlo7@alumnes.ub.edu; elicomelles@ub.edu

Phraseological studies (Howarth, 1996 & 1998; Granger, 1998; Moon, 1998; Hunston & Francis, 2000; Wray, 2002; Oakey, 2002a/b; Biber, 2006; Hyland, 2008; Granger & Meunier, 2008, among others) have pointed to the fact that phrases are central to the grammar. In this context, phrases are viewed as the main focus of linguistic analysis since they contribute to characterise the language of different discourse communities.

Research on the field of phraseology indicates that mastering formulaic constructions is a significant factor in successful linguistic production. Gaining control of the phraseological conventions of different genres is what makes writers be considered members of a given discourse community. Writing accurately in a foreign language is already a difficult task. It becomes even more difficult when the ultimate goal is not only producing error-free texts but also mastering the prototypical conventions of a specific genre, which is essential to the effective production of texts. According to Gledhill (2000a; b), while scientists usually know how to use specialised vocabulary, non-natives do not show a good command of general terms prototypical of the scientific genre. In addition, he has also shown that lack of phraseological knowledge is characteristic not only of non-native but also of junior native scientists.

As English is now the international language of academic communication, it is important for those who are are committed to disseminating their research internationally to show phraseological competence when writing in this language. Consequently, the creation of useful analytic reference tools that improve non-native scientists’ usage of multiword units in specialised English appears to be a fruitful area of research.

Such a need is confirmed by the inclusion of a section 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, this phraseological approach has not yet been used in more specialised dictionaries. Although existing technical and scientific monolingual dictionaries provide terminological and encyclopaedic information or, in the case of bilingual and multilingual dictionaries, they offer translation equivalents, they usually lack contextual information on the syntactic and collocational patterns of non-technical vocabulary used in English medical texts. This may lead to unsatisfactory look-up experiences, especially when one needs information about the context on which the meaning of a given lexical entry depends. It is thus necessary to develop lexical databases and specialised dictionaries that take into account the lexico-grammatical patterning of lexical items that can help to enhance users’ performance.

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 in relation to the sequences they are typically found in. The information displayed by SciE-Lex in its earlier stages includes the following: Word class, Morphological variants, Equivalent(s) in Spanish, Patterns of occurrence, List of collocates as well as Examples of real use and explanatory notes to clarify usage.

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 will be supplemented with 3- to 5-word clusters (e.g. consistent with the/these/previous results, based on the/these results, these/our results suggest/indicate/demonstrate that), and will provide explicit information about their frequency, composition and function. Formulaic constructions will be classified according to their function in the discourse. The new additions in SciE-Lex will 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 casuistry.

References


(pp. 85-105). Bern: Peter Lang.
In a recent book, Tarp (2008: 121) suggests using the term *leximat* to identify lexicographical reference works based on modern information technology. He defines a leximat as “a lexicographical tool consisting of a search engine with access to a database and/or the internet, enabling users with a specific type of communicative or cognitive need to gain access via active or passive searching to lexicographical data, from which they can extract the type of information required to cover their specific needs”.

In this contribution, we will show why traditional, paper-based electronic dictionaries do not meet these requirements as well as how such a new kind of reference work could then be conceived. The (new interface of the) *Base lexicale du français* (BLF), which is an online learner’s dictionary (or lexicographical database) for intermediate and advanced learners of French, will be used as an example of a leximat.

Traditional, paper-based electronic dictionaries have improved the access to data by adding powerful search functions and/or integrating several resources, for instance on a single CD-ROM. On the *Collins-Cobuild Resource Pack On Cd-Rom*, for example, you can find the *Collins COBUILD Advanced Learner’s English Dictionary* and the *Collins COBUILD English Grammar* as well as a thesaurus or text samples. However, they often fail to quickly provide the answer to simple but frequent questions asked by language learners, such as the correct spelling of irregular word and verb forms (Pruvost, 2003). The consultation of these electronic dictionaries also requires from the learner almost the same dictionary reference skills as for paper dictionaries, skills which are, as is commonly acknowledged, often inadequate. Miller (2006: 441) even calls it “the perennial problem of lack of dictionary skills and reluctance to use a dictionary”.

In an electronic dictionary the large amount of information on words is organised in a well-structured way. Therefore, it should be possible to develop a more user-friendly electronic dictionary which allows a more efficient way of accessing the specific information which the learner needs, thus meeting his real needs and answering his questions. Hausmann (1977: 144) already listed a few of these questions more than thirty years ago; Tarp (2008: 77) completes this list of needs.

We used both lists as a starting point to create a new kind of user-oriented interface for our online electronic dictionary for learners of French, the *Base lexicale du français*. We identified five basic needs:

- get information on: the user has a word (form) in a foreign language in mind and wants more information about it (gender, spelling, morphology, meaning, lexical profile, etc);
- get the translation of: the user wants the translation of a (multi)word (expression) from his mother tongue into a foreign language;
- verify: the user wants to check if a word (combination) is used in a
language or if the translation he has in mind is correct;

- learn: the user wants to understand how a language is structured, how he can avoid common errors;
- make exercises: the user wants to practise his knowledge of the language.

Once the user has identified the kind of basic need he has, he can make a pop-up screen appear by moving the cursor to each of the links on the homepage. In this pop-up screen there is a textbox for the word or word combination the user wants to receive information on. Alternatively, he has to click on one of the GO!-buttons in order to obtain the precise information he wants, for instance the gender of the word *problème*, the translation of *au detriment de* in German, the meaning of the word *dessein* or a list of all adjectives which express the idea of *intensification* in French.

The learner is led directly to the requested information without seeing a common dictionary entry.

In order to access all this information, which is only partly available in our database, convenient shortcuts to a number of external websites are integrated in the *Base lexicale du français*, such as links to the *Opus* parallel corpus website for the translation to and from various languages (Tiedemann & Nygaard, 2004). In this way, the *Base lexicale du français* becomes a special kind of single web portal to the most relevant websites devoted to French vocabulary.

A complete tracking and logging system records all traffic on the website by identifying more than 250 paths which the users can follow through the database and the external websites. This is far more than similar recordings by De Schryver et Joffe (2004) and Měchura (2008), which concentrate on the words entered on the homepage of online dictionaries. When we have enough records of the use of our online dictionary, we hope to get a better insight into the way people use (electronic) dictionaries. These observations could then, in turn, be used to improve the interface. Another advantage of our interface is that it can easily be extended to other languages, a project we are currently working on.

The new interface of the *Base lexicale du français* is an attempt to create a more user-friendly interface for a lexicographical tool which better meets the user’s needs. It also shows how to gain maximum benefit from the integration of various online lexical resources.

**References**


**Websites**

BLF (Base lexicale du français): ilt.kuleuven.be/blf

Collins-Cobuild Resource Pack On Cd-Rom:


Opus – an open source parallel corpus: http://urd.let.rug.nl/tiedeman/OPUS/
Lexicography in the Grid Environment

Marta Villegas, Nuria Bel, Santiago Bel, Francesca Alemany & Héctor Martínez
Universitat Pompeu Fabra
Marta.villegas@upf.edu, Nuria.bel@upf.edu, Santiago.bel@upf.edu, francesca.alemany@gmail.com, hector.martinez@upf.edu

The research reported in this paper is part of the activities carried out within the CLARIN project. CLARIN is a large-scale pan-European project devoted to the creation of a persistent and stable network serving the needs of the European Humanities and Social Sciences research community. The CLARIN infrastructure strongly relies on the SOA (Service Oriented Architecture) approach to bring together the range of resources and tools available in the research community and make them accessible to others.

New methods enable new research by giving researchers access to distributed resources, which include data collections, computing resources, scientific tools and visualisation. All these resources constitute a distributed computing network or Grid.

In a Grid environment, researches have access to Web Services which allow them to execute services on a remote system. In this new scenario, researches 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 and workflow edition and orchestration.

This paper reports our experience when e-lexicography moves into a grid environment in order to integrate different tools. This allows evaluating the move to grid-lexicography focusing on the requirements, interoperability problems and workflow edition, what will help in demonstrating the feasibility of benefiting from grid technology in the lexicographic research.

In order to test and validate our research, we defined and implemented a real use case. Our scenario is how to give support to lexicographers who are creating a dictionary for Spanish as a second language (DAELE). The work is developed with the support of a Lexical Markup Framework (ISO 24613) compliant lexicographical platform, COLDIC.

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 corpus in order to find out usage and co-occurrence patterns. Often they have to struggle with a large number of occurrences for a particular word. Analysing these large and unstructured lists becomes a highly consuming task. The objective, therefore, is to provide lexicographers with clustering processes that automatically structure occurrence lists in a sensitive manner.

Clustering of KWICs concordances is performed on two different bases: lexically based and pattern based.

In the first case, the system provides a representation based on the bag of words model in order to eventually classify the occurrences according to the other
words occurring in the concordance. Thus, the system calculates different similarity measures based on the bag of words method (that is, based on the co-occurrence of words in different samples). These calculations are performed amongst the whole set of occurrences retrieved. Similarity measures are then collected in vectors, so that each vector represents the similarity between a pair of occurrences. Finally, these vectors are sent to the clusterer, which will consider the data in each vector to form clusters of occurrences according to their lexical similarity.

In the second case, the system checks the syntactic characteristics of the contexts of each occurrence and builds the corresponding vector. The so built vectors are then grouped using clustering techniques, so that the lexicographer can ask for one or more examples of the clusters found. Again, the system structures a large set of examples in a sensitive manner.

The scenario, therefore, requires the integration of different tools in a single workflow:

1. the dictionary tool (COLDIC),
2. the access corpus & concordances extractor
3. vectorization process and
4. WEKA for clustering.

In order to achieve the integration of these tools in a single process, we made them available via Web Services, thus we provided Web Service wrappers around these applications.

Once all required processes were deployed as Web Services, we were able to use a workflow editor to define and execute our workflow. In this case, we used the Taverna workbench, a free software tool for designing and executing workflows, created by the myGrid project and used by researchers from the bioinformatics field.

Essentially, the points which deserved more attention were: (format) interoperability and orchestration.

- (Format) interoperability: Typically, a workflow consists of a series of processors which from a given input eventually return a final output. Processors require input(s) in a given format and return output(s) in a specific format. WSDL (Web Service Description Language) types inputs/outputs using basic types such as string or integer. This type system is clearly not enough and new typing methods are required. In our case, we integrate different tools in a single workflow, this means that inputs & outputs need to be compatible (for instance, we need that the output of the KWIC tool be the input of the vectorization process and that the output of this process can be loaded into WEKA). Here we suggest for extensive use of valid xml for input/output and the corresponding Schema definition, so that required format conversion can be addressed via XSLT (also implemented as a transformer Web Service).

- Orchestration: This apparently trivial scenario involved different tasks which need to be managed and arranged in a certain specific way. Thus, when defining a workflow, the researcher needs to specify these arrangements. In this case, workflow orchestration needs to include: conditional control structures (such as 'if then' conditions), ‘for each’
loops and iteration. In addition, some of the process require huge amount of processing. This is the case of the vectorization process, where for each occurrence example the system compares it against the rest of the samples in order to generate the corresponding vectors. This task can be very time & process consuming and a parallel processing proves to be very useful.

References


KEPLER project web site [http://kepler-project.org](http://kepler-project.org).


myGRID project web site, [http://www.mygrid.org.uk/](http://www.mygrid.org.uk/)


W3C web site, [http://www.w3.org/](http://www.w3.org/)

Using Aligned Corpora to Construct Large-Scaled Linguistic Resources: The Electronic Dictionary of Spanish Compound Nouns

Duško Vitas¹, Takuya Nakamura² & Stavroula Voyatzí²
¹University of Belgrade, ²Université Paris-Est Marne-la-Vallée
vitas@matf.bg.ac.yu; nakamura, voyatzí@univ-mlv.fr

Parallel corpora have proven useful for a variety of purposes, including but not limited to statistical machine translation (SMT) (Brown et al., 1993; Véronis, 2000) and "annotation projection" across languages as training data for NLP tools of various kinds (Yarowsky et al., 2001; Hwa et al., 2002; Mitkov & Barbu, 2004; Pado & Lapata, 2005). Moreover, if the corpora's annotations include manual word alignment information, they can be used, e.g. for evaluating word alignment approaches (Mihalcea & Pedersen, 2003; 2005).

Relatively little work has been made on systematic hand annotation of word alignment information in parallel corpora, i.e. on explicitly marking correspondences at the word level. These annotations have often been ad hoc and restricted to few language pairs; furthermore, they rarely include complex linguistic information like multiword units (MWUs). Moreover, the size of parallel and multiparallel corpora that offer larger sets of word alignments is rather small.

On the other hand, monolingual resources for the majority of European (and not only European) languages, which allow a considerably more complex processing of corpora than in the case of statistically based methods, have been developed within the RELEX network¹. These resources² are freely available for research and business with the LGPLLR license. The integration of an Open Source corpus processor such as Unitex³ (Paumier, 2006) into the exploitation of multilingual resources is particularly useful (Vitas et al., 2006). The Unitex software has the option for the alignment of two texts as well as the concordancer that enables the necessary correction of alignment to be performed and queries to be posed over the parallel aligned text (bitext). For all texts that form a bitext, Unitex produces during preprocessing the corresponding dictionaries of simple words, compounds and unrecognized words (Paumier & Dumitriu, 2008). Such an enhancement of the Unitex system offers a new challenge: transferring linguistic resources from one language to another. In fact, Unitex enables to use hand annotations of MWUs incorporated in the source monolingual corpora and align them in the target monolingual corpora which lack such information. This is particularly interesting for the NLP community as it facilitates considerably the construction of new linguistic resources (Laporte, to appear).

In this paper, we explore the possibilities to construct an electronic dictionary of compound nouns in Spanish using an existing one in French, and by means of parallel corpora aligned by a tool compatible with Unitex. Our motivation derives from the intuition that in corpora of two reasonably similar languages, i.e. languages

¹ http://infolingu.univ-mlv.fr/Relex/Relex.html.
of a same "family" like Romance, compound nouns identified in the source language are likely to correspond to compound nouns in the target language as well. In order to verify the pertinence of this idea, we applied our method to a sample of the French-Spanish parallel corpus of the European Parliament (EUROPARL; Koehn, 2005). The method we propose here consists of three basic steps: parallel corpus alignment, monolingual MWUs annotation, transfer of annotations from one language to another.

The system we used to align Europarl is XAlign, an Open Source text alignment tool developed at Loria (Bonhomme & Romary, 1995). Two factors determined this choice: its compatibility with Unitex and its ability to realize one to one correspondences in two texts which allowed us to identify equivalent segments in the French-Spanish parallel corpus. Manual control was performed during automatic alignment in order to detect and mark eventual missing parts in both corpora, which happens frequently in Europarl corpus. Bitexts obtained with this method are first aligned at the sentence level (marked by <seg> tags) in different text formats (TEI, TMX, html) and by means of the WS4LR software tool (Krstev et al., 2006). The actual size of the corpus is approximately 500,000 words per language, i.e. one million words of French-Spanish bitext. It may be visualised as a TEI document with Unitex.

In order to annotate the French corpus, we tagged all the occurrences of compound nouns described in a morphosyntactic lexicon (Courtois & Silberztein, 1990), following the same principles and method as (Laporte et al., 2008); we revised the annotation manually. We also annotated proper names and two classes of named entities: temporal expressions and numerical expressions (Martineau et al., 2007). Three types of features are included in the annotations: internal morphosyntactic structure (e.g. NDN stands for the Noun - Preposition de - Noun structure), inflectional features (e.g. ms stands for masculine singular) and semantic type in the case of proper nouns and named entities (e.g. Toponym refers to place names, TIMEX denotes a temporal entity). A sample of the annotated French corpus is given below:

<seg>Comme vous le savez , le {NE type="TIMEX"}matin du 21 septembre{/NE} , un terrible {N cat="NDN" fs="ms"}tremblement de terre{/N} , d’amplitude de 7,6 sur l’{NP cat="NDN" fs="fs" type="Mesure"}échelle de Richter{/NP} , s’est produit sur l’{NP cat="NDN" fs="fs" type="Toponym"}île de Taïwan{/NP} .</seg>

We used a program of utility that enabled to transfer semi-automatically the tags on segments in French onto Spanish texts. At the end of this process, we obtained a bitext with correspondence information regarding various types of MWUs:

<seg>Como saben ustedes , en la {NE type="TIMEX"}mañana del 21 de septiembre{/NE} , se produjo en {NP}Taiwan{/NP} un terrible {N}terremoto{/N} de 7,6 grados en {N}la escala Richter{/N} .</seg>

1 http://poincare.matf.bg.ac.yu/~vitav/ParallelTextTagEditor(v3).rar.
2 At this stage, morphosyntactic and inflectional information is only available for French MWUs, but it is also envisaged to provide such information for the corresponding Spanish MWUs either during the transferring process or afterwards.
We also obtained the list of corresponding MWUs in French and Spanish which constitutes a useful database for the study of Spanish vocabulary, notably to define lemmas or structures of compound nouns.

Our goal is to extract a sufficiently large number of Spanish MWUs in order to make available for the NLP community free large-scaled linguistic resources in Spanish. The software and method described here will be of interest to researches with diverse backgrounds in natural language processing as they combine statistical measures of co-occurrence with knowledge-lite modules of word categorization, morphological variation and MWUs recognition. The complete results and the evaluation of our method will be published in the final paper.

References


In *Proceedings of the HLT/EMNLP*, Vancouver, Canada.
Building an Electronic Combinatory Dictionary as a Writing Aid Tool for Researchers in Biology

Alexandra Volanschi & Natalie Kübler
UFR EILA, University Paris Diderot – Paris 7
Alexandra.volanschi@eila.univ-paris-diderot.fr, nkubler@eila.univ-paris-diderot.fr

The present paper reports on a method of exploring the combinatorial properties of terms belonging to a specific field of biology, yeast biology, based on the analysis of a corpus of scientific articles. This research has led to the production of a writing aid tool meant to help non-native authors write scientific papers in English. The tool meets the needs of young French researchers, who are constrained to publish in English as early as the Post-doctoral level. The imperative which governs a researcher’s career nowadays, “Publish in English or perish”, is rendered discouraging both by the lack of specialised dictionaries and by the lack of teaching materials targeted for these needs.

In order to better estimate our users’ needs we sent out a questionnaire to the teaching and research members of the Life Sciences Department at the University Paris Diderot. The results analysis has shown that almost 96% of all scientific publications are written directly in English and that 90% of participants to the questionnaire use other scientific articles as a writing aid. The kind of information they search in pre-published articles is - to the same extent – scientific information and hints on phraseological information such as obligatory prepositions, connectors, terminological collocations (to clone, express, cut, carry a gene) but also collocations belonging to “general scientific language” (Pecman, 2004), such as to strengthen, reinforce, support a hypothesis. Mastering phraseological information is one of the elements proving a scientist’s belonging to a scientific community.

In order to extract terminological collocations specific to yeast biology but also collocations belonging to general scientific language we built a specialised corpus, composed of research articles on yeast biology, selected with the help of biologists working at the University Paris Diderot. We have thus gathered a large working corpus of over 5.5 million words, which we have POS tagged and parsed using the Stanford dependency parser (Marneffe, 2006).

In the first research stage (reported in this article) we focused on restrictive collocations, for which we supplied the following working definition (by adopting a number of defining features discussed – among others – in Hausmann (1989), Benson (1986) or Lin (1998)): restrictive collocations are recurrent binary combinations, the members of which are in a direct syntactic relation. As the orientation of the collocation (between the base and the collocative) is parallel to that of the syntactic dependency, we adopted a hybrid automatic collocation extraction method similar to that of Lin (1998) or Kilgarriff & Tugwell (2001).

The hybrid collocation extraction method we devised is based on the dependency parsing of our corpus. We first extract co-occurrences of items in a given syntactic relation. Unlike the methods cited above, we do not pre-define the syntactic patterns we are interested in, but rather eliminate a number of auxiliary relations (such as negation, or determination, although they should be subject to further investigation) and examine all remaining syntactic relations. The method uses a common association
measure, mutual information, in order to sort co-occurrences extracted on the basis of syntactic patterns, and a few extra heuristics (frequency and coverage) in order to distinguish collocations from free combinations and one author's idiosyncrasy.

Using this hybrid method we extracted collocations occurring at least three times in the corpus, in at least three different documents, recording, at the same time, the frequency of occurrence, the number of documents in which they appeared and the mutual information of the co-occurrence. Results of this extraction process were included in an electronic dictionary the preliminary version of which may be consulted online at the address:
http://ytat2.ijm.jussieu.fr/LangYeast/LangYeast_index.html.

Choosing an electronic dictionary format has allowed us to use both bases and collocates as entries in the dictionary and supply one illustrating example for each candidate collocation. The preliminary version of the tool contains the combinatory profiles for 2810 nouns, 1034 verbs and 1334 adjectives, containing more than 78 000 collocations.

Several improvements of the dictionary may be envisaged. Among other things a lexicographical validation of the dictionary entries (selected mostly on frequency criteria), supplying more illustrating examples for each entry, and – most importantly – finding a way of presenting results better adapted for the end users of our dictionary, for whom notions such as “regisseur”, “argument” or “modification nominale” are irrelevant.

The writing aid tool we wish to supply for biologists writing in English as a second language will be extended in two research directions which we have begun to explore. On the one hand we wish to extend our analysis to the argumental structure of a number of specialised verbs taking into account syntagmatic constraints on verb arguments. These structures, also extracted from the dependency parses of the corpus by analysing all dependency relations related to the verb, should provide biologists with a clearer picture on the verb usage. Collocational analysis can only provide a partial picture of this. On the other hand, we envisage extending our analysis form restricted collocations (which we define as binary recurrent combinations) to larger collocational complexes (cf. Howarth, 1996) or usage patterns (such as idiomatic formulae specific to scientific discourse).

Finally, we envisage using the dictionary we have developed as well as the corpus from which it is derived in English for Special Purposes courses for biologists and building teaching materials derived from these resources.

References


Dialect Dictionaries at a Crossroads: Multiple Access Routes on the Example of the Dictionary of Bavarian Dialects in Austria (Wörterbuch der bairischen Mundarten in Österreich (WBÖ))

Eveline Wandl-Voigt
Institut für Österreichische Dialekt- und Namenlexika (IDINAMLEX) / Institute for the Lexicography of Austrian Dialects and Names; Österreichische Akademie der Wissenschaften / Academy of Sciences
eveline.wandl-vogt@oeaw.ac.at

Dialect dictionaries are traditionally long term projects, usually not very open to modernisation and changes due to the fact that every change in long term projects means the investment of a lot of money.

Yet, some European projects in the past time managed the challenges, so the project WBÖ in Vienna, published since 1963 by the nowadays Institute for the Lexicography of Austrian Dialects and Names (Institut für Österreichische Dialekt- und Namenlexika (IDINAMLEX)). The working group established a database in 1993 to store the dictionaries base material. 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 opened to the world for the first time web-based, georeferenced and interactive.

In 1998 a rationalisation concept was issued to the WBÖ targeting the completion of the dictionary in 2020 as a (virtual) unit consisting of the printed dictionary and the complementary database. This so called Straffungskonzept was altering the dictionaries structure effectively. New types of entries have been established (so called Datenbankartikel [database entry]) The mediostructure of the dictionary changed.

(1) Example: Simple database entry (historical base material):
†Diaun, Gerichtsbote obVintschg. (16.Jh.), s. DBÖ

(2) Example: Simple database entry (recent base material):
Trikó,, M., N., Trikot, elast. Stoff; best. Eng anliegendes Kleidungsstück ugs., s. DBÖ

The WBÖs access structure is problematic (e.g. due to the macrostructure itself, the etymological-historical headword, the highly sophisticated structure of the entry itself) and neither really functional nor user-friendly at all.

(3) Example: The standard German equivalent Apfelbaum (‘apple tree’) corresponds with the WBÖ-headword (Apfel)pāum and (Epfel)pāum which itself is a subentry of the WBÖ-main-entry Pāum.

---

1 Further information about e-lexicography at the institute compare Wandl-Voigt (2008b).
4 WBÖ 5,27.
5 WBÖ 5,512.
6 Example entries compare WBÖ-Beiheft 2 pp.14-17.
7 WBÖ 2,621.
Due to this very specific headword-tradition getting the WBÖ into digital surroundings and linking its content with other dictionaries and databases means effort.

Within the project Database of Bavarian dialects in Austria electronically mapped (Datenbank der bairischen Mundarten in Österreich electronically mapped (dbo@ema)) multiple access routes have been developed since 2007.4

Several different access routes will be presented:

First, I will focus on the (interactive, web-based) map as navigation tool for the dictionary and database content as well, e.g. headword, base-material, WBÖ-entries, furthermore bibliography and (lexicographical) documentation. This suits users who are often interested in material originating from a certain location or area.

Second, a concept of phonetic access5 on the example of the WBÖ should be discussed. Problems of reducing phonetic navigation into practice are heterogeneous data (notation systems of about 2,000 collectors and co-workers).

Independently, it seems to be a scientific and technical challenge to enable phonetic navigation and visualize phonetic realization for a dialect dictionary to increase user friendliness step by step.

References


---

1 WBÖ 5,731: a in the headword(s) with dot above (not to be realized in this font) signalizes not German etymology in the WBO.
2 WBÖ 5,753.
3 WBÖ 5,767.
4 Scholz (2008), Wandl-Voigt (2008a); wboe.at (19.06.2009).
5 Ryš (2009); Sobkowiak (1994).
6 WBÖ 5,440.
7 WBÖ 5,115.
Sobkowiak, W. (1994). Beyond the year 2000: Phonetic-access dictionaries (with word-frequency information). In EFL.
Integration of Multilingual Terminology Database and Multilingual Parallel Corpus

Miran Željko
Government of the Republic of Slovenia, Translation division
miran.zeljko@gov.si

Electronic dictionaries, glossaries and terminology databases are usually still made to be used like printed dictionaries: the user has fairly limited context in terms of usage and (s)he can check only one word at a time. We have tried to overcome these limitations in our terminology database.

In theory, the general idea that dictionary and corpus should be somehow integrated was stated by Wofgang Teubert (1999: 312); he defined it more precisely later (Teubert, 2004: 17).

The combination of monolingual dictionary and monolingual corpus has been available for several years – on the web, for instance, in the Digital Dictionary of the German Language (http://www.dwds.de/). Bilingual dictionaries (and terminology databases) are even more useful for translators – but it is more difficult to make a bilingual or multilingual dictionary and even more difficult to make a multilingual corpus of useful size. An example of bilingual dictionary and corpus was made by Tomaz Erjavec (1999), but it was not developed into a useful system.

To make things easier, we started with a terminology database instead of a dictionary. Terminologists in the Slovenian government started to compile a multilingual terminology database (Evroterm) during the EU accession period and today it contains about 100,000 terms. During the EU accession period, we started to compile a bilingual (English-Slovene) corpus of translations (Evrokorpus), which now contains more than 60 million words. However, the real counterpart to the multilingual terminology database is a multilingual corpus. It was possible to make a multilingual corpus when the European Commission's Directorate-General for Translation made publicly accessible its multilingual translation memory for the Acquis Communautaire (http://langtech.jrc.it/DGT-TM.html). We thus created Termacor (terminology and corpus) software that combines a multilingual terminology database and multilingual parallel corpus.

Termacor's unique features are:

- one user interface for terminology and/or corpus search (from a translator's point of view, a corpus is just a logical extension of the terminology database)
- the user can select one source language and any number (up to 22) of target languages
- it is possible to see basic or detailed data on a particular term from the terminology database or a particular segment from the corpus database
- results from the terminology search are highlighted in the corpus output – the user can thus find his/her point of interest faster
- links are provided to full texts of documents.

Another useful tool for terminologists and translators is our Terminator software for terminology analysis (http://evroterm.gov.si/x/indexe.html). This software analyses a text supplied by the user and transforms the terms found in the text into hypertext links that provide information stored in the Evroterm/Evrokorpus databases. Terminator can be used in several ways:

- When a translator gets a text that has to be translated, (s)he can easily see which terms are stored in the terminology database and thus the terminology is more consistent in the translated texts. This is especially important if several translators translate texts from the same field.
- When a terminologist receives a new-term table from a translator, (s)he first analyses this table with the Terminator. In this way, (s)he can easily see which terms are really new and which already exist in the database and may only need correction.
- When a terminologist checks an existing text for possible terminology candidates that could be added to the terminology database, the candidates can be recognised much faster than by reading normal text.

The software and its databases are subject to continuous development, so by the time of the eLexicography conference, these databases will contain additional data and the software will have additional features.

References


Reverse Access via an Index Based on the Notion of Association.
What Do Vector-Based Approaches Have to Offer?

Michael Zock¹ & Tonio Wandmacher²
¹Laboratoire d’Informatique Fondamentale
CNRS, UMR 6166, F-13 288 Marseille
²Institut für Kognitionswissenschaft
Universität Osnabrück, Germany
michael.zock@lif.univ-mrs.fr, tonio.wandmacher@uni-osnabrueck.de

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 are keen 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. More precisely, we would like to enhance an existing electronic dictionary to help people find the word they are looking for even in the case of partial or imperfect input.

One of the most vexing problems authors encounter is their failure to access in due time a word they are certain to know. This is generally known as the tip-of-the tongue problem (Burke et al., 1991). That authors know the word, i.e. that they have memorized it, can often be shown, as quite so often they end up producing it later, while they failed a moment ago. In the case of word-finding problems, people tend to reach for a dictionary, which does not guarantee, of course, that they will find what they are looking for. There are various reasons for this, some of them being on the lexicographers’ side: (a) it is not easy to anticipate the various kinds of user inputs; (b) in what terms shall these inputs be couched (primitives), etc.

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 (BBI, OECD), Bernstein’s Reverse Dictionary and Rundell’s MEDAL (2002), 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 psychologists have shown (Brown et McNeill, 1966), speakers experiencing word finding problems know generally many things about the lexeme they are looking for: parts of the definition, etymology, beginning/ending of the word, number of syllables, part of speech etc. We would like to use this information, no matter how poor it may be, to help the authors to find the word they are looking for. In other words, given some input we will try to guide their navigation, providing hints to
lead them towards the target word.

To achieve this goal we will build on an idea described in Zock & Schwab (2008), who 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 to capture the target user’s world knowledge and build, metaphorically speaking, 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’).

Search is then quite straightforward. The user provides as input all the words coming to his/her mind when thinking of a given idea or 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.

Again, there remains 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 & Schwab (2008) use direct co-occurrence measures (1st order approaches) to determine association, there is some evidence that 2nd order approaches, based on co-occurrence vectors, are more suited to this end. Word space models like LSA (Latent Semantic Analysis) or HAL (Hyperspace Analogue to Language) represent each term of a given vocabulary as a high-dimensional vector, calculated from co-occurrence in a large training corpus. This allows to determine semantic relatedness on the basis of the distance of the respective vectors which can now be calculated.

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 & Levy, 2007). This is a very important point, as word finding problems generally occur with low frequency terms. For this reason simple, but broad-coverage approaches, like the web-based methods applied by Sitbon et al. (2008) could turn out to be more appropriate for our purpose.

The goal of this work is to shed some light on the advantages and disadvantages of vector-based approaches as opposed to 1st order association measures with regards to lexical access, i.e. finding words. As this is work in progress, we cannot present a thorough evaluation yet, but we plan to test several of the measures mentioned above on the TOT data set generated by Burke et al. (1991). This should allow us not only to show what kind of associations is most relevant for lexical access, but also to reveal the particular strengths and weaknesses of the different measures. As an ideal outcome these insights should enable us to generate an optimal association matrix, generated from a combination of several singular measures and techniques.

References

Boissière, P. (1862). *Dictionnaire analogique de la langue française : répertoire*