Violeta Seretan and Eric Wehrli. Context-sensitive look-up in electronic dictionaries. In Rufus H. Gouws, Ulrich Heid, Wolfgang Schweickard, and Herbert E. Wiegand, editors, Dictionaries. An International Encyclopedia of Lexicography. Supplementary volume: Recent Developments with Focus on Electronic and Computational Lexicography, volume 5/4 of Handbooks of Linguistics and Communication Science (HSK), pages 1046–1052. De Gruyter Mouton, Berlin, Boston, 2013.

71. Context-sensitive look-up in electronic dictionaries

1. Introduction

The advent of the digital era and of Internet drastically changed the way in which dictionary information is accessed and presented. The tedious look up of words in traditional paper dictionaries has been replaced by the ubiquitous, instant access to electronic dictionaries containing additional multimedia elements (like pronunciation, images, and videos).

Most of the existing electronic dictionaries were conceived as simple copies of paper dictionaries in machine-readable format, and their entries can be accessed by a headword-based search. In the past decade, however, a new generation of electronic dictionaries emerged, that propose a more sophisticated look up methodology. Instead of requiring users to enter a citation form, they aim to take advantage of the developments in the language technology field in order to provide searching capabilities that are more tailored to the actual users' needs.

Provided that users are likely to consult the dictionary while reading a text, it is apparent that the inflected word form together with its context should play a major role in accessing the dictionary information.

Context-sensitive dictionaries perform a linguistic analysis of the word and its context in order to enable and to refine the look up process by linking the inflected word form to its base form listed in dictionaries, and by narrowing the information to be displayed according to the clues provided by the word context.

This chapter introduces the problematics of the context-based dictionary look up, presents the available context-sensitive dictionaries, and indicates the development perspectives for this look up approach.

2. Context-sensitive dictionary look up

Simple context sensitivity is achieved when the look up procedure of an electronic dictionary integrates a lemmatization module for retrieving the base form (lemma) of the searched word. A word that is unknown to the user is also likely to have a morphological structure that is difficult to decode, especially in agglutinative languages. The morphological analysis is therefore a necessary step in finding a match in the dictionary entry list, which usually displays base word forms rather than inflected forms. Often, the lemmatization has to be complemented by part-of-speech (POS) tagging that disambiguates the lexical category, as a word form can belong to different categories (for instance, the word *bear* can be either a noun or a verb). The major challenges faced by the look up procedures of this kind are the recognition of lemma variants, as well as the definition of a mapping between the tagging and lexicographic POS categories.

These procedures simplify the users' task in relating the unknown word to the appropriate main entry and to the subentry corresponding to the compatible lexical category. An example of simple context-sensitive tool is GLOSSER-RuG, described in the next section. But more elaborated context-sensitive look up methods exist that are aimed at a finer selection, which is in fact possible with the current state of the art in automatic language analysis. Dictionary entries may be dissuasively long, and displaying them entirely is of little help for the user. The methods based on context analysis aim to select only those subparts of the identified entry (or subentry) that are compatible with the word context, and to hide the irrelevant information. In accordance with the observation that the meaning of a word is determined by its context (cf. the Firthian slogan "You shall know a word by the company it keeps!" (Firth 1957, 179)), they take into account the linguistic environment of the searched word in the task of discriminating between the possible interpretations provided in the dictionary.

First of all, the context analysis carried out by these methods has to consider the possibility that the searched word is part of a multi-word unit (henceforth, MWU). MWUs (e.g., compound nouns, phrasal verbs, idioms, collocations) are prevalent in language; they make up a high proportion of the information stored in dictionaries, in which they appear as headwords of subentries of single

words, as headwords of main entries, or can be found among the usage examples provided. The success of a look up procedure depends on the matching algorithm employed for determining if the word context contains one of the MWUs listed in dictionary. For certain MWUs (such as compounds), a simple string comparison would suffice; however, a high number of MWUs allow for a certain amount of syntactic freedom that challenges their recognition in text. Therefore, the solution generally adopted is to syntactically analyse the context sentence. Sometimes, more than one match is possible between text and dictionary MWUs, and the look up procedure has to define strategies for choosing what results to display; then, certain heuristics and criteria (e.g., the longest-match or the closest-match criterion) are used for ranking competing candidates.

Secondly, regardless of the fact than the looked-up word is part of a MWU or not, context-based look up procedures may try to operate a sense selection for the ambiguous words. Usually, dictionaries contain additional information for guiding the user in selecting the appropriate sense of a word, and this information is exploitable by automatic look up procedures. Such information includes domain labels (e.g., *Scol*, indicating 'school'), indicator fields (e.g., *in sport*), or typical collocates (like *burden*, for the verb *to bear*). Only a few look up procedures have been implemented that take into account this information in their attempt to discriminate between multiple senses. They rely, for instance, on the semantic tagging of the word context, but their portability across languages is at present hampered by the reduced availability of semantic analysis tools.

In addition to the challenges that are inherent to the look up methodology proper, the design of context-sensitive dictionary look up engines must also comply with particular user-interface requirements. The ideal comprehension aid application is fast, unobtrusive, available from any application or from any part of the computer screen, gives access to user dictionaries as well, displays both the selected subentry and (optionally) the whole entry, and provides a means for storing the displayed results.

Typically, a context-sensitive look up engine must tackle the following issues (the focus of a specific engine may change according to its desired usage):

(i) capturing the input: detecting the word that the user wants to look up, retrieving the text surrounding it (e.g., by taking into account the text selection or the mouse position), and, additionally, isolating the sentence containing that word;
(ii) decoding the input: performing the morpho-syntactic analysis of the context sentence in order to obtain the lemma for the searched word and to recognize the MWUs it may be part of;
(iii) entry match: relating the obtained lemma or MWU with a headword of a main dictionary entry (in the second case, the success of the match should not be precluded by the specific position of the searched word within the MWU);
(iv) subentry match: choosing the entry subpart that is most compatible with the interpretation of the word in context (i.e., the specific POS category, the subsuming MWU, and the appropriate reading);
(v) results presentation: choosing a screen location and a format for displaying the matched information, providing access to extended information, and, optionally, re-inflecting the output according to the morphological analysis of the input.

3. Existing context-sensitive dictionaries

3.1. COMPASS/LOCOLEX

COMPASS (described, inter alia, in Feldweg/Breidt 1996 and Breidt/Feldweg 1997) was a project carried out jointly by industrial and academic research centers (among which Rank Xerox Research Centre Grenoble, Fraunhofer Gesellschaft Stuttgart, University of Tubingen, University of Lyon, and Bournemouth University). The project took place between 1994 and 1996 and aimed at building a tool for comprehension assistance intended for humans reading texts in a foreign language (http://www.sfs.uni-tuebingen.de/Compass/). LOCOLEX (patented by Xerox) is the dictionary look up module on which COMPASS is based, and it relies, in turn, on the IDAREX formalism for describing MWUs in a local grammar framework, as shown in Breidt/Segond/Valetto (1996).

The language directions considered were German−English and English−French. The two dictionaries used (CGE and, respectively, OH) were first converted in the LOCOLEX internal dictionary data format, in which MWUs are reduced to their canonical form and are represented as IDAREX regular expressions (IDAREX stands for 'idiom de-

scription as regular expressions'). The canonical form of a MWU contains only the lexically-fixed components in the correct ordering, as well as meta-expressions like *sb*, *sth*, *one's*, *oneself*. Besides, the dictionary data was validated and augmented by corpus-based lexicographic revision.

In order to allow the recognition of MWUs in text, the dictionary was also augmented with local grammars in the form of finite-state networks specifying the valid morpho-syntactic environment for these units. Thus, a detailed description in terms of the allowed morphosyntactic manipulation was provided for several thousands of MWUs in English, French, and German. Such variation include, for instance, changes of inflection, addition of modifying words, and syntactic re-structuring of components in phenomena like passivization, scrambling, and separable prefix verbs.

LOCOLEX controls the actual look up process, by checking the compatibility between a MWU in the (identified) dictionary entry and the sentence context. The system itself is language-independent, but relies on language-specific text analysis tools which it applies on the text side. In particular, it uses the Xerox finite-state tools for building a finite-state network for the input sentence, which it then compares against the finite-state network in the MWU's description. If the two match, then LOCOLEX marks the corresponding dictionary subentry for presentation in the user interface.

The user interface of the prototype consists of a text editor supporting HTML text that allows three interaction modes: modification of text, annotation with translation notes, and comprehension assistance. When the last mode is activated, a pop-up window displays, for the selected word, the translation found in the matched dictionary subentry. Optionally, the whole dictionary entry can also be displayed. The window disappears after a preset time of inactivity.

The screen capture of this interface shown in Breidt/Feldweg (1997) provides an example of context-sensitive translation involving the sentence 'Auch 1993 schreibt Lufthansa rote Zahlen', with *schreibt* as the selected word. The output proposed, *to be (deeply) in the red*, is the translation of the whole unit the word is part of (*rote Zahlen schreiben*), whereas the non-contextual translation is *to write*.

The usage examples provided in dictionaries are not taken into account in the matching procedure, since considered important only from the language production perspective. Further development plans included the use of shallow parsing in order to improve the recognition of verbal MWUs, especially in languages with a relatively free word order (like German), as well as the use of dictionary meta-information (collocation lists, domain labels) and of word sense disambiguation techniques in order to select the appropriate reading for a word, depending on its context.

3.2. GLOSSER-RuG

GLOSSER-RuG – presented, among others, in Nerbonne/Dokter (1999) – was a prototype developed in the 1990s at University of Groningen in the framework of a research project involving different other partners (Rank Xerox Research Centre Grenoble, University of Tartu, The Bulgarian Academy of Science, and MorphoLogic Budapest). The goal of the project was to support the reading of texts in a foreign language. GLOSSER-RuG targeted, in particular, Dutch students reading French texts.

The system performs the morphological analysis of words and provides access to dictionary entries on the basis of POS-disambiguated lemmas. In addition, it displays corpus examples containing the selected word or its inflectional variants. The morphological analysis and POS disambiguation in context is carried out using Xerox's finite-state technology. Context-sensitive dictionary look up is partly achieved for fixed, contiguous MWUs (like *idée fixe*, *'obsession'*) by checking the presence of all of these items in the text.

The user interface consists of a main window displaying the source text, and of three auxiliary windows showing the information specific to the selected word: the morphological analysis, the dictionary entry, and corpus examples. A Web demo is currently available online at http://www.let.rug.nl/glosser/Glosser/.

3.3. DEFI Matcher

DEFI Matcher – see, for instance, Michiels (1998) and Michiels (2000) – is a sophisticated program for matching text and dictionary entries that was developed at University of Liège in the framework of the DEFI project carried out between 1995 and 2000 (http://engdep1.philo.ulg.ac.be/michiels/matcher.htm). The ultimate goal of this project was

to build an online comprehension tool that guides the human readers in selecting the most appropriate translation in context. Since the program was not intended as an end-user product, most of the efforts concentrated on the matching procedure rather than on its user interface.

DEFI Matcher made use of both bilingual and monolingual dictionaries (OH and RCEF bilinguals, LDOCE, COBUILD, and CIDE monolinguals), as well as of additional resources like word thesauri (ROGET, WORDNET). The language pair dealt with was English−French for which both directions were initially considered; however, due to the lesser availability of linguistic tools for French, the French−English direction was less developed than the other.

The program was implemented in Prolog (its source code was made available online). Also, the dictionary information was converted into Prolog internal database format. The prototype developed can be used in text mode, by supplying both the context sentence and the word to translate, and by specifying the desired match mode (single-word lexeme mode or multi-word unit mode). Sample inputs are shown below.

(i) single-word lexeme mode: ability s > I admire his ability to solve problems.
(ii) multi-word unit mode: bear m > The point doesn't bear any relation to the question we are discussing.

The output of the matcher is a list of translations ranked according to the compatibility with the textual form. Partial results for the query in example (ii) above are:

(1) 181 − 281206, [bear],
to bear no relation to,
tr(n'avoir aucun rapport avec,
être sans rapport avec)
(2) 168 − 281177, [bear],
to bear a relation to,
tr(avoir rapport à)
(3) 113 − 7401, [bear],
I can't bear his preaching to me,
tr(je ne supporte pas qu'il me
fasse la morale)
(4) 87 − 281220, [bear],
to bear some resemblance to,
tr(ressembler à, offrir une
ressemblance avec)
(5) 68 − 281211, [bear],
to bear oneself,
tr(se comporter)

DEFI Matcher relies on the partial syntactic analysis performed on both the text and dictionary entry side with the LingSoft *engcg* surface parser for English. The rough analysis provided by *engcg* is further enhanced by *tagtxt*, an in-house tool that builds on the top of its output by recognising noun phrases, hypothesising syntactic relations for chunks, and computing voice and polarity features.

The syntactic analysis enables the mapping between the dictionary information and the input text, and is argued necessary for those MWUs not occurring in text in their canonical form, as well as for expressions with lexicographic fillers (*something*, *someone*). This matching approach is less constrained in comparison with local grammar approaches recording the exact amount of syntactic manipulation and lexical variation admitted by a given MWU.

The direction of match is from dictionary to text. The match procedure attempts to find, iteratively, a mapping for each of the items of a dictionary MWU in the sentence context. Each item has to be mapped, but on the text side elements can be skipped or can be found in a different order.

A characteristic of the matcher is that it uses dictionary examples, which are treated just like MWUs. Their specific voice and polarity features are checked for compatibility with the text, and preference is given to the closest match: the closer the MWU to the text, the better. The underlying rationale is that lexicographers have good reasons to list a MWU in a form that deviates from the canonical form and, thus, to record the passive voice (e.g., *we've been had*) or the negation (*no earthly reason*). Yet, multiple translations could be pertinent, and it is argued that the appropriate answer is not a binary decision, but a continuum. A pertinence weight is assigned to competing translations based on the partial weights computed for each item in the MWU, as well as on a global weight depending on whether the unit has a phrasal or a sentential status (sentential units are likely to be examples).

The latest developments in DEFI were centered towards using dictionary metalinguistic information (such as the indicator field and the list of collocates), word thesauri, as well as definitions and examples from monolingual dictionaries for improving the match procedure.

For instance, in a sentence context like 'And now that little bogy has been exorcised',

the verb *exorcised* provides strong disambiguation clues for the noun *bogy* for which multiple translations are possible: *croquemitaine* (the right choice in this context), *bogée*, *spectre* and *croute de nez*. In order to promote *croquemitaine* in the top of the result list, the system applies a complex procedure that consists of: parsing the sentence in order to recover the collocate bearer (i.e., the verb exorcise); looking at its typical collocates listed in the dictionary (e.g., *demon*, *memory*, *past*); and trying to establish a link between these collocates and the disambiguating words provided in the indicator field of bogy (namely, *in nose*, *in golf*, *frightening*, *evil spirit*, *imagined fear*, *bugbear*). The selection of the right sense succeeds when the link between *demon* and *evil spirit* (the indicator of *croquemitaine*) is established based on information from monolingual dictionaries.

3.4. MoBiDic/MoBiMouse

MoBiDic/MoBiMouse, presented in Prószéky (1998) and Prószéky/Kis (2002), is a commercial tool developed by MorphoLogic (http://www.morphologic.hu/). It provides context-sensitive translations for words by matching the surrounding of the input text against dictionary information (including usage examples).

The match is based on shallow parsing that attempts to identify MWU candidates in the context of translation point. The parsing is performed with a syntactic analyzer based on HUMOR, a morphological analysis tool available for a number of highly-inflectional languages (e.g., Hungarian, Polish, Czech, and Romanian).

If the parser fails to recognize a MWU, the match is done simply on the basis of word stems. It is considered as successful when all the content words from the dictionary side are also found in text around the translation point, in a window of configurable length. An example of dictionary search was provided by the authors for the word *dog* found in the context *to lead a dog's life*. The results obtained include:

(1) dog {dog, dogs, dog's, dogs'} − 21 occurrences in expressions of the basic dictionary,
(2) dog AND life − 2 occurrences in expressions of the basic dictionary,
(3) lead AND dog − 1 occurrence in expressions of the basic dictionary,
(4) lead a dog's life − 1 occurrence as an expression in the basic dictionary.

The matched dictionary subentries are presented in a formatted way to the user. In addition to text, the output can contain multimedia elements, such as pictures and pronunciation. Multiple dictionaries can be looked up at the same time, including user dictionaries in XML format. Currently, the largest available dictionaries are English−Hungarian and German−Hungarian.

Unobtrusiveness is argued as the most distinguishing feature of this tool. There is no separate window for the application, which is activated when the mouse hovers over a word on the screen. The input text is captured with accurate OCR technology. The matched translation is displayed in a bubble-shaped pop-up window, which disappears when the mouse is moved again. Also, since the dictionaries distributed are compact, they can be copied on the computer hard disk when the application is installed, so that the CD-ROM drive remains available to the user.

3.5. Sharp

Poznanski et al. (1998) describe a practical glosser built at Sharp Laboratories of Europe Ltd., in which an English text is annotated with Japanese translations from a bilingual dictionary. The dictionary look up is based on lemmatization and POS disambiguation of the English text. In addition to single words, (possibly discontinuous) MWUs are detected as interruptible n-grams and are looked-up in the dictionary, e.g., *stem … from*. A pop-up window displays the translation alternatives. In the input text, the source MWUs are underlined and marked with different indexes and colors in order to facilitate their identification by the user.

3.6. Benedict

Benedict – The New Intelligent Dictionary, introduced in Löfberg et al. (2004), was a joint academic and commercial project carried out between 2002 and 2005 by Kielikone Ltd., University of Lancaster, University of Tampere, Nokia, and the publishing houses HarperCollins and Gummerus (http://mot.kielikone.fi/benedict/). It aimed at building a context-sensitive look up tool able to select the appropriate sense of a word in English and Finnish monolingual dictionaries.

The sense selection method relies on the semantic analysis of the word's context pro-

vided by USAS, a semantic tagger developed at the University of Lancaster (cf. Rayson et al. 2004). After a mapping is defined between the tags employed by USAS and the lexicographic domain taxonomies, a matching algorithm (DDS, the Domain Detection System) computes the semantic distance between the context of the word and the dictionary definitions. The relevant main entry is preliminarily selected based on morphological analysis and on simple phrase detection.

Prototypes and demonstrators for this match engine have been developed for Windows, Pocket PC, and Nokia's Symbian operating system, and have been commercially released as part of the *MOT Professional* dictionary software. This system features advanced capabilities of word search in the structured dictionary information or in a corpus (including search using wildcards and Boolean operators, as well as fuzzy search of near matches), and offers a KWIC functionality for displaying results.

In the separated desktop version, Benedict software is activated when the user clicks on a word on the screen. The surrounding text is retrieved using OCR and is semantically tagged, then the word form is looked up in the dictionary; finally, the most compatible sense found identified by DDS is highlighted in the pop-up window showing the dictionary entry.

Benedict is capable, for instance, to retrieve the right sense for the word *arm* (as "an object that covers or supports the human arm, esp. the sleeve of a garment or the side of a chair, sofa, etc.") when it occurs in a sentence like 'He was sewing arms onto the jacket'.

3.7. TWiC

TWiC – Translation of Words in Context (described in Wehrli 2003 and Wehrli 2004) is a system for online terminological help developed since 2003 at the Language Technology Laboratory of the University of Geneva in connection with the ITS-2 translation system. It accesses the bilingual lexical databases used by this system, and relies on the syntactic analysis of the word context provided by the parser Fips – see, for instance, Wehrli (2007) – in order to identify the correct lexical entry as well as the MWU the word may be part of. Accordingly, it proposes a translation for both the word and MWU, which is compatible with the linguistic environment.

TWiC supports a high number of language pairs, as the parser is available for several languages, i.e., French, English, German, Italian, and Spanish (other languages, like Greek, are partly supported as well). However, most of the bilingual lexical information currently available concerns the pairs with French as the source or the target language.

The detailed morphosyntactic analysis provided by the parser Fips enables TWiC not only to discriminate between ambiguous lexical categories (for instance, the word *rose* can be a noun, an adjective or a verb, whereas in a context like 'They woke up before the sun rose' the verbal reading is the only possible), but also to identify the MWU the selected word may be part of (e.g., *rose garden*, or *to rise up*). In this case, TWiC displays the translation of the whole unit as well (i.e., for the examples cited, *roseraie* and *se soulever*).

The most distinguishing feature of TWiC stands in its ability to recognize the numerous MWUs (notably, the verb-object collocations) whose items are not necessarily adjacent in text since they may undergo various grammatical operations, like relativization, passivization, interrogation. The recognition of these units cannot be done at a superficial level because the object is extraposed, as illustrated in the examples below containing the collocation break-record:

(1) Records are made to be broken.
(2) There are several records that may not be broken.
(3) Yet another world record has been broken by one of our members.
(4) Which record will he break first?

This collocation is successfully identified thanks to the deep syntactic analysis provided by Fips. In the sentence normalization returned by the parser, the extraposed elements are linked to their canonical argument positions; therefore, the noun record is connected to the canonical position of object via the antecedent-trace chains created by Fips. Thus, TWiC is able to match the obtained canonical form with the corresponding lexical entry, and to return the translation for the whole MWU (*battre un record*).

Corpus-based collocation extraction and translation techniques, such as those described in Seretan/Wehrli (2006) and Seretan/Wehrli (2007), are employed in order to populate the bilingual lexical databases of the system.

TWiC is available as a plug-in for two major Internet browsers, downloadable from http://www.latl.unige.ch/. The application (which is activated with a mouse click) first retrieves the sentence containing the selected word and identifies the source language by using a language guesser; then, it analyses the sentence and displays the following results in a window: the base form of the reading that is compatible with the syntactic context; its translations; and the translation of the subsuming MWU, if any. The window can also be used for specifying a different source and target language.

An extension of this application is *TwicPen*, the offline version of TWiC that can be used on printed documents thanks to a scan-held scanner device that digitalizes the input text (cf. Wehrli 2006).

4. Conclusion

Several context-sensitive dictionary look up engines have been developed so far in the framework of academic or commercial projects, or as joint work. Even if most of them did not evolve beyond the prototype stage and are currently inactive, they paved the way to a new generation of 'intelligent' electronic dictionaries that is emerging in our multilingual information society. Their efficiency is subject to the advances in the field of natural language processing, from which they borrow the language analysis tools used for finding the best match between the word context and dictionary information. While state-of-the-art technology relies on deep syntactic parsing or on semantic tagging, further improvements could be achieved by combining the syntactic and semantic techniques, by employing word sense disambiguation methods, as well as by harnessing the existing lexical semantics resources.

5. Selected bibliography

Breidt, E./Feldweg, H. (1997): Accessing foreign languages with COMPASS. In: Machine Translation 12, 153–174.

Breidt, E./Segond, F./Valetto, G. (1996): Formal description of multi-word lexemes with the finite-state formalism IDAREX. In: Proceedings of COLING, 1036−1040.

CGE = Collins German−English English−German Dictionary. Glasgow 1991.

CIDE = Cambridge International Dictionary of English. Cambridge 1995.

COBUILD = Collins Cobuild English Dictionary. London and Glasgow 1987.

Feldweg, H./Breidt, E. (1996): COMPASS: An intelligent dictionary system for reading text in a foreign language. In: Papers in Computational Lexicography 53–62.

Firth, J. (1957): Papers in Linguistics 1934−1951. Oxford.

LDOCE = Longman Dictionary of Contemporary English. Harlow 1978.

Löfberg, L./Juntunen, J.-P./Nykanen, A./Varantola, K./Rayson, P./Archer, D. (2004): Using a semantic tagger as dictionary search tool. In: Proceedings of EURALEX, 127–134.

Michiels, A. (1998): The DEFI matcher. In: Proceedings of EURALEX, 203–211.

Michiels, A. (2000): New developments in the DEFI Matcher. In: International Journal of Lexicography 13, 151–167.

Nerbonne, J./Dokter, D. (1999): An intelligent word-based language learning assistant. In: Traitement Automatique des Langages 40, 125–142.

OH = Oxford-Hachette French Dictionary. Oxford 1994.

Poznanski, V./Whitelock, P./ IJdens, J./Corley, S. (1998): Practical glossing by prioritised tiling. In: Proceedings of ACL-COLING, 1060–1066.

Prószéky, G./Kis, B. (2002): Context-sensitive electronic dictionaries. In: Proceedings of COLING, 1−5.

Prószéky, G. (1998): An intelligent multi-dictionary environment. In: Proceedings of ACL-COLING, 1067–1071.

Rayson, P./Archer, D./Piao, S./McEnery, T. (2004): The UCREL semantic analysis system. In: Proceedings of the LREC workshop on Beyond Named Entity Recognition − Semantic labelling for NLP tasks, 7−12.

RCEF = Collins-Robert English/French Dictionary. Glasgow 1995.

ROGET = Roget's thesaurus, public domain version downloadable from various websites.

Seretan, V./Wehrli, E. (2006): Accurate collocation extraction using a multilingual parser. In: Proceedings of COLING/ACL 2006, 953–960.

Seretan, V./Wehrli, E. (2007): Collocation translation based on sentence alignment and parsing. In: Proceedings of TALN 2007, 401–410.

Wehrli, E. (2003): Translation of words in context. In: Proceedings of Machine Translation Summit IX, 502–504.

Wehrli, E. (2004): Traduction, traduction de mots, traduction de phrases. In: Proceedings of TALN, 483–491.

Wehrli, E. (2006): TwicPen: Hand-held scanner and translation software for non-native readers. In: Proceedings of COLING/ACL Interactive Presentation Sessions, 61–64.

Wehrli, E. (2007): Fips, a "deep" linguistic multilingual parser. In: Proceedings of ACL Workshop on Deep Linguistic Processing, 120–127.

WORDNET = WordNet Prolog Package, downloadable from the Princeton University Website: http://www.cogsci.princeton.edu/~wn/.

Violeta Seretan, Eric Wehrli,
Genève (Switzerland)

72. Large-scale documentary dictionaries on the Internet

1. Introduction

By large-scale documentary dictionaries we understand dictionaries aimed at or designed to cover large macrostructures and, at the same time, a substantial amount of detail in their microstructure. Furthermore, in order to be represented in this article, dictionaries must be either available through the World Wide Web (interchangeably used with the term Internet in this article) or at least attempt a web presence in the production phase of their compilation. A description of all dictionaries worldwide meeting these two criteria would by far exceed the scope of this article. Therefore we impose additional restrictions:

- We only consider monolingual reference dictionaries covering the general language with some encyclopedic and cultural material; learner dictionaries as well as interlingual dictionaries are not taken into account.
- We focus on large national projects. Dictionaries sharing the above-mentioned characteristics are all long-term projects compiled either by large national institutions or less frequently by well renowned private publishing houses.
- We consider only the European context, and here in particular we focus on the six languages of the large countries of the European Community (EU-G6, Group of Six), namely English, French, German, Italian, Polish and Spanish. In addition, we describe three projects in other EU-countries that are remarkable either because of their lexicographic and technical approach or because of their relevance for the linguistic identity of these nations, namely those dealing with Danish, Dutch and Hungarian.

This article gives an overview of such projects, in terms of their basic facts as well as their commonalities and differences with respect to underlying guidelines and principles, both lexicographic and managerial respectively issues of access.

1.1. Lexicographic aspects

- Does the project draw on existing print dictionaries or is the dictionary born digital, i.e. compiled exclusively on a digital basis? This question correlates with the following one: does the dictionary follow the linear, print-oriented order or is it organized as a lexical database where e.g. systematic linking between elements or onomasiological dependencies can be expressed?
- Does the project compile a synchronic or a historical dictionary? Dictionaries generally do not draw a sharp line between both alternatives. Thus, dictionaries qualified as synchronic may well contain diachronic elements such as etymological remarks or a chronology of citations. However, the description of entries is always related to a given language stage or a time interval. On the other hand, dictionaries qualified as historical are based on historical principles, i.e. they document the changes in form and meaning of words starting with their first appearance in the language.
- Does the project rely exclusively on electronic text corpora or does it use a mixture of corpora and paper slips?
- Which number of entries is attained or attempted by the project? Even though it is difficult to compare the size of the dictionaries on the basis of their entry numbers alone – e.g. to which extent are derived entries and compounds counted as entries? – it still is an meaningful indicator of the dictionary size. Related to the size of the macrostructure is the question of the fine-gradedness of the microstructure: it is be-