Gregory Grefenstette, Rank Xerox Research Centre
Ulrich Heid, Universität Stuttgart
Bruno Maximilian Schulze, Universität Stuttgart
Thierry Fontenelle, Université de Liège
Claire Gerardy, Université de Liège

The DECIDE Project: Multilingual Collocation Extraction

Abstract

One of the problems facing the user of a bilingual dictionary is producing multiword expressions and phrases in the target language when the explicit phrasal translation does not appear in the dictionary. Defining collocations as the preferred choice of words for expressing the desired concept, the DECIDE project has been exploring how collocational information from mono- and bilingual dictionaries and raw text corpora can be discovered, extracted and stored online. During the project, we have developed tools for identifying potential collocations from raw text; for marking up English, French and German text for use in an interactive corpus query tool; for accessing lexical and grammatical patterns over such a corpus via this corpus query tool; for accessing collocations derived from online bilingual dictionaries; and for documenting such collocations using available text corpora. Finally, we have produced a common interface to these textual, corpus and dictionary tools, and used this interface to create a multilingual lexicon of the collocational choices of support verbs for nominalizations of speech act verbs. This paper presents an overview of this European Union sponsored project, its objectives, its methodology, and its results.

1. Introduction

One of the major problems facing the user of a bilingual dictionary is producing multiword expressions and phrases in the target language when the explicit phrasal translation does not appear in the dictionary. The DECIDE[1] project has been producing a response to this problem involving the extraction of collocational information from both multilingual dictionaries and raw text corpora. Information from both sources is stored in machine-accessible multilingual lexicons, so that the results may be exploited by traditional dictionary builders as well as by automated language processing. In this paper, we will describe the motivations for this project, its objectives, the methodology used to attain these objectives, and the results produced.

2. Motivation

In an influential paper entitled "On collocations", Mackin (1978) tackles the problem of how to teach collocations to foreign language learners. He stresses the fundamental distinction between production and understanding and argues that collocations do not pose any serious problems in the understanding process. Any non-native speaker is likely to recognize and understand a collocation but using collocations and selecting the appropriate term is much more difficult and may even be considered as one of the most serious stumbling blocks in language learning. This is also why Mackin argues for the compilation of specialized dictionaries, since it is generally admitted that collocations cannot be accounted for in terms of grammatical rules. They are an element of our lexical knowledge. Their idiosyncratic nature makes them particularly well-suited for inclusion in a special type of dictionary designed not so much for decoding, i.e. understanding, text (like most traditional dictionaries) as for encoding text.

In a more recent paper, Cowie argues that "journalistic prose draws very heavily on verb-noun collocations that are already well-established and widely known" (Cowie, 1992:1). This does not seem to be typical of journalistic prose only, however, and research in applied linguistics has shown that native speakers often memorize ready-made word combinations. They usually have some predisposition to store these combinations as wholes, which accounts for the pervasiveness of lexical collocations in everyday language. The term 'lexical collocation' can be used here to refer to the privileged, idiosyncratic relationship that holds between some verbs and their subjects/objects or between some nouns and the adjectives that modify them. This notion of lexical collocation is illustrated by the example of a bilingual dictionary user who wishes to encode a concept involving two or more words in the target language. The encoder may be sure of the translation of one word in the expression, but the other words have to be chosen from a number of likely candidates. Knowing the proper lexical collocation means choosing the candidate a native speaker word choose. Examples of encoding French concepts into English collocations are *faire attention* translated as *pay attention*, *faire tes devoirs* as *do your homework*, *mettre la table* as *set the table*, and so on. In these cases, the encoder might well be certain of which noun to use and uncertain about the corresponding verb.

3. Objectives

The question of privileged relations led the DECIDE consortium to consider whether a lexical collocation dictionary might be automatically constructed from textual sources where these idiosyncracies may be implicitly contained. In bilingual dictionaries, some collocational information is given but often only the superordinate term of a collocating item is indicated because exhaustive listing of all the hyponyms would be too space-consuming. In large text corpora, lexical collocations should give rise to detectably higher cooccurrence of their items. We decided to examine how these resources could be mined for collocational information, and how the result of this mining could be made useful for a bilingual encoder.

We defined the aim of the DECIDE project as threefold: (i) To carry out a state-of-the-art survey and in-depth assessment of existing corpus-oriented and dictionary-oriented tools for the extraction of collocations, (ii) to design and implement a toolbox for the extraction of such information, with special emphasis on monolingual corpora and bilingual dictionaries, (iii) to prove the feasibility and the relevance of extracting and combining collocational information from corpora and dictionaries by elaborating prototype collocation lexicons for a restricted subset of vocabulary. The principal objective of DECIDE was to be able to unite collocational information coming from both raw text corpora and bilingual dictionaries into a human and machine usable multilingual lexicon.

4. Methodology

The methodology adopted was to first study the existing tools that could be applied to the problem. Most major European dictionary publishers were contacted but only one (Van Dale) agreed to grant free access to a copy of the software they were using or developing. Most British publishers agreed to let us have access to their products by inviting us to visit them. One of the advantages of these visits was that we were able to outline the general features of what lexicographers actually expected from collocation extraction tools. The consortium partners also acquired and tested a variety of text analysis and corpus query tools. There were two types of tools: various tools developed in universities and the tools used by publishing houses. We produced surveys (Fontenelle et al. 1994a, Schulze et al. 1994, Schulze et al. 1994b) of these tools for the European Commission[2] which contained a general presentation of the

main functions and parameters that lexicographers would like their tools to offer and a comparative table summarizing the specifications (platform, text format, size of corpora, etc.) and functionalities (statistical measures, display, filtering and sorting of information, etc.) of each tool studied.

After this initial study, the collocation extraction problem was attacked in three different directions: (i) A lexical database was defined to house collocation information extracted from type-setting tapes of a bilingual dictionary. This database was then constructed and filled (Fontenelle et al. 1994b, Fontenelle et al. 1995); (ii) Extensions to lexicographical tools (Schulze and Christ 1994) for marked-up corpus examination and query were designed and developed (iii) Specifications of tools for collocation extraction from raw text and for marking-up raw text were defined (Grefenstette et al, 1994), and then implemented by the partners (Grefenstette et al. 1995).

As for raw lexicographical material on which to test these tools, we chose a set of more than one hundred speech act nouns for the languages: English, French and German; existing corpora of many million words in each language, available to each partner; and several online bilingual dictionaries. We set ourselves the task of using these tools to derive a multilingual lexicon of support verbs, defined in a wide sense (Mel'cuk 1984), for these nouns.

5. Project Developments

5.1 Interactive Lexical Semantic Database

A machine readable version of the *Collins-Robert English-French Dictionary* (Atkins and Duval 1987) was transformed[3] into a lexical-semantic database, in which each entry was enriched with Lexical Semantic Functions, according the Meaning-Text-Model of Mel'cuk (1984). Fontenelle (1995) details the extraction of the collocational elements from the bilingual dictionary, the manual disambiguation of collocational clues and the manual addition of Lexical Functions, as well as the development of a program, called *robcol*, for accessing the database.

In order to give the user maximal freedom as far as manipulation of files is concerned, we decided to create a command line interface to *robcol* in the DECIDE project. Examples of this interface are the following:

robcol -i milk gives all the instances of the records containing *milk* as an italicized word (collocate) in the *Robert-Collins.*

*robcol -h "*tion"* gives all the records where the headword ends in *tion.*

robcol -i "fox"*

*-pos n -lex "*son*"* returns all the records where the italicized word starts with *fox*, where the headword is a noun, and where the Mel'cukian lexical function name includes the string *son.*

One of the possible uses of the program is to reconstruct the semantic network around a given base, which appears in italics under different entries in the printed dictionary. If we are interested in finding out the headwords which contain the string *fox* in italics, we may type the following command: *robcol -i fox*, which yields the following results, giving the user a more complete vision of the semantic network around the concept:

bark (n) : ~fox~ => glapissement <m> (renard,s0son)
bark (vi) : ~fox~ => glapir (renard,son)
bitch (n) : ~fox~ => renarde <f> (renard,female)
brush (n) : ~fox~ => queue <f> (renard,part)
...
run (vt) : ~fox~ => chasser (renard,)
stink out (vt sep) : ~fox~ => enfumer (renard,)
yelp (n) : ~fox~ => glapissement <m> (renard,s0son)
yelping (adj) : ~fox~ => glapissant (renard,a0son)
yelping (n) : ~fox~ => glapissement <m> (renard,s0son)

The output format here is: *headword (part of speech) : italicized item => French translation of the headword (French translation of the italicized item* [here : renard] + *the standard lexical function or lexical-semantic relationship [s0son*: noun expressing the typical sound; *sloc*: noun for the typical place of something]). As can be noticed, the notion of lexical function (Mel'cuk 1984) has been extended to include other types of semantic relations such as the part-whole relation (noted as *part*).

The program makes it possible to retrieve collocational information by combining different types of criteria. In the following query, for example, the user is interested in extracting the verbs which can take *law* as direct object and whose meaning with respect to *law* can be expressed in terms of the lexical function *liqu* (liquidate, eradicate):

robcol -i law -lex liqu

abolish (vt) : ~law~ => abroger (loi,liqu)
annul (vt) : ~law~ => abroger (loi,liqu)
do away with (vt fus) : ~law~ => supprimer (loi,liqu)
repeal (vt) : ~law~ => abroger (loi,liqu)
rescind (vt) : ~law~ => abroger (loi,liqu)
revoke (vt) : ~law~ => rapporter (loi,liqu)

It should be noted that the output of such a query makes it clear that it is possible to extract highly complex collocations with phrasal-prepositional verbs such *as do away with*. Another example of a query combining lexical functions and the base of a collocation is the following: *robcol -i liar -lex magn* which can be paraphrased as follows: list items expressing a 'high degree or intensity' (*Magn* in Mel'cuk's terminology) of the noun *liar*. This command yields the following adjectival intensifying collocates: *arrant*, *bare*, *chronic*, *confirmed*, *habitual*, *hopeless*, *out-and-out*, *rank*, *straight and unqualified*, with their accompanying French equivalents.

The program 'robcol' can of course also be used to retrieve information as if one were simply looking up a headword in the printed dictionary. The primary access key is therefore the headword field and the information which is displayed is similar to what can be found in the published version of the dictionary, although the output is reformatted. The interest of such information in a generation perspective should be self-evident. For example, *robcol -h confirmed -pos adj* shows that the *drunkards, liars, smokers, bachelors, sinners* and *habits* can all be *confirmed.*

If one is interested in retrieving the subset of support verbs which includes inchoative verbs that may collocate with the noun *habit*, one may type the following command:

robcol -i habit -lex incepoper1

The lexical function *incep* denotes the beginning of a process and is here combined with support verb function *oper1* to form what Mel'cuk calls a complex lexical function. The Collins-Robert data for this query is the following :

acquire (vt) : ~habit~ => prendre (habitude,incepoper1)
contract (vt) : ~habit~ => prendre (habitude,incepoper1)
develop (vt) : ~habit~ => contracter (habitude,incepoper1)
form (vt) : ~habit~ => contracter (habitude,incepoper1)
take to (vt fus) : ~habit~ => prendre (habitude,incepoper1)

In addition to the integrating the implicit collocational information from bilingual dictionaries into a user queryable form, the DECIDE project also modified existing tools for extracting collocational information for marked and unmarked text. The next two sections describe these modifications.

5.2 Corpus Query Tools

One of the aims of DECIDE was to find ways of extracting information about lexical collocations so that the language encoder can know which among a number of lexical choices are more likely to be used by a native speaker. A review of German dictionaries for the verbs collocating with, for example,*Vorschlag*, shows that there is considerable variation between what is proposed to the user, as well as what might be found in a corpus (last column):

	Dictionaries				
Vorschlag	Wa	La	Ag	Du	Corpus
einen V. machen	+	+	+	+	+
einen V. annehmen	+	+	+		+
einen V. ablehnen	+		+	+	+
einen V. zurückweisen	+			+	
einen V. akzeptieren			+	+	+
einen V. verwerfen			+	+	+
einen V. ignorieren			+		+
einen V. übergehen			+		+
einen V. mißachten			+		+
etw in V. bringen	+		+		+
sich zu einem V. aufraffen			+		+
sich zu einem V. entschließen		+		+	
einen V. billigen			+		
sich einen V. erlauben			+		
sich jds V. anschliessen			+		
auf jds. V. eingehen			+		

In order to find collocations in marked-up text, the corpus query tool CQP (Schulze and Christ 1994) has been considerably extended so that macro-procedures corresponding to patterns can be defined. For example, the *German Verb 1st-order* macro isolates all the verbs appear-ing before*Vorschlag* throughout the same sentence in a part-of-speech tagged corpus, by generating the following corpus query:

[pos = "V."][]* [word = "Vorschlag" pos = "N.*"] within s;*

The output of this query can be directly viewed as:

..nn , ist ungewiß . Der Senat *änderte den Beschluß Vorschlag* , daß die Rostocker Hi
..n Straßen . Kein Verständnis *äußerte Hobrack für den Vorschlag* , die Clara-Zetkin-
..rn noch in Bagdad aufhielt , *äußerte Unterstützung für den Vorschlag* und sagte Peki
..? "Staatsrat Hoppensack *überbrückte den Ressort-Konflikt mit dem Vorschlag* , die
..nn-Pohl (CDU) zusammen und *übergab ihr einen Vorschlag* des Westberliner
..er wieder packen . Der Senat *überläßt die politischen Vorschläge* bislang Heinrich
..nlösen . Senat und Sparkasse *überlegen Vorschläge* , wie mit den bekannten
..Informationen der *taz übernimmt der Antrag des Landesvorstandes viele Vorschläge*
..rie des Grünen-An trags : Er *übernimmt inhaltlich voll den Vorschlag* der Initative
..Beim Studierendentag der GEW *überrascht FU-Professor mit Vorschlag* zur
..Sadtentwicklung , Schreyer , *überrascht von des Bausenators Vorschlägen* zum
..Herwig Haase *überraschte am Dienstag die Öffentlichkeit mit dem Vorschlag* , am

Further macro-procedures allow the results of this query search to be tabularized, passing the results through lemmatisers, and producing output such as the following, giving the verbs collocating with *Vorschlag* over a 100 million word German corpus:

Frequency

machen	140
kommen	63
unterstützen	53
geben	53
halten	51
begrüßen	41
nennen	37
finden	36
bezeichnen	34
stoßen	32
folgen	31
sehen	29
reagieren	29
gehen	29
begründen	24
stimmen	22
unterbreiten	21

5.3 Corpus Extraction Tools

Raw text exploitation tools (morphological analyzers, part-of-speech taggers, and low-level parsers) were extended and modified to extract collocations and to mark-up English or French input text automatically so that it could be exploited by the DECIDE corpus query tools (section 5.4). The text tools took raw text as input, apply morphological analyzers, and statistical part-of-speech taggers. The tagged text was then parsed using a low-level parser (Grefenstette 1994). Extensions to the parser extracted information about verb and noun syntactic patterns and automatically marked-up the text for use in the corpus query package. In the following Helsinki-style markup *<s>* *<np>* and *<vp>* mark beginnings of sentences, noun chains and verbal chains. The first column gives the surface form, the second column gives the part of speech, the third gives the lemma, and the fourth indicates low-level syntactic relations.

<s>			
<np>			
correlation	*NOUN*	*correlation*	*NN>*
coefficients	*NOUN*	*coefficient*	*DOBJ>*
</np>			
<vp>			
have	*INF*	*have*	*AUX*
been	*BE*	*been*	*AUX*
determined	*PPART*	*determine*	*MAINV*
</vp>			
<np>			
between	*PREP*	*between*	*PREP>*
the	*DET*	*the*	*DET>*
levels	*NOUN*	*level*	*<IOBJ*

Potential subcategorization information was extracted by the parser which makes choices as to what words are in sytactic relations to other words. A recent evaluation of SEXTANT (Grefenstette, 1996) over a technical manual corpus showed that it was able to recall 81% of such relations with a precision rate of 83%.

Frequency	**Base Noun**	**Collocating verb-SYNTACTIC RELATION**
14	*effect*	*have-DOBJ*
7	*effect*	*study-DOBJ*
6	*effect*	*produce-DOBJ*
5	*effect*	*exert-DOBJ*
5	*effect*	*determine-DOBJ*
3	*effect*	*see-DOBJ*
3	*effect*	*note-DOBJ*
3	*effect*	*enhance-DOBJ*
3	*effect*	*demonstrate-DOBJ*

From this information automatic corpus queries which feed into the CQP system are generated, which look like this

```
( meet ([lemma="effect" & syntag="DOBJ.*"],
    [lemma="have" & syntag="MAIN.*" ] s ) expand to s;
```

which would generate the KWIC lines of all the sentences where *have* had *effect* as a direct object.

Currently versions of SEXTANT are being produced for European languages in addition to English and French at the Rank Xerox Research Centre in France.

5.4 Common Toolbox Interface

A common interface has been developed at Liège that pilots and interfaces information derived from the online dictionaries, the raw corpus data and the interactive corpus query tools. The main components of the toolbox are: The Collins-Robert En-Fr/Fr-En database, described in section 5.1, the corpus markup tools for French and English developed by Rank Xerox, described in section 5.3, and the corpus exploitation tools developed at the University of Stuttgart, of which certain extensions such as the macro-preprocessor are described in section 5.2. A menu-based language described in Appendix 3 of Gerardy et al. (1996) allows any of these tools to be called and executed.

6. Resulting Multilingual Lexicon

The first decision to be made by the consortium about the lexicon was whether the collocational information contained there should be restricted to support verbs stricto sensu (*make, take* etc.) corresponding to the Mel'cukian *oper1* class of verbs, or whether it should be extended to verbs collocating regularly with speech act nouns (such as *proffer an apology, harbour suspicion*, etc.). Since the latter category of verbs is most interesting for NLP systems and non-native speakers of English, it was decided to include this wider category of verbs in the lexicons. A trilingual list of speech act nouns based on Wierzbicka (1987) was produced by the consortium members to serve as the basis for the lexicon.

The construction of the lexicons followed two different schemes, one for German, another one for English and French. This asymmetry is due to the fact that for English and French, lexical information from the Robert/Collins English-French dictionary, as well as from other monolingual and bilingual dictionaries was available, whereas only little information on German collocations could be extracted from machine-readable dictionaries. The two different architectures converge however, in so far as, following the basic principles and layout of the DECIDE project, information from text corpora and information typically available in dictionaries were combined. The basic inventory of information includes nouns, the support verbs they can be combined with, whenever possible, a lexical function in terms of Mel'cuk's descriptive model, and, for all entries, information on the relative corpus frequency of the collocation and its syntactic behaviour.

For English and French, all this information except the frequency information could be extracted from dictionaries and, comparatively, from corpora. Frequency could, of course, only be measured in corpora and was thus added to the other information available, which led to a process in which a validation of the information retrieval from the dictionaries by comparison with the corpus material took place. For German, however, only a basic set of information from dictionaries was available and more information had to be extracted from corpora by using slightly more complicated analysis templates designed to extract syntactic and frequency information from the text material.

7. Conclusion

This project has successfully integrated information coming from three different sources: (1) Online bilingual dictionaries, which were augmented with Lexical Function information for all collocations, (2) Raw text markup which prepared text for use in the corpus query system, as well as extracted frequency information for various configurations of support verbs over the subset of speech-act nominalizations (for English and French), (3) And from corpus evidence (for German) derived through automated querying and treatment of marked-up text. The lexicons produced for a set of one hundred speech-act nominalizations has been made available online(2) as an illustration that the corpus and dictionary mining techniques produced here can provide useful information about collocation constructions.

8. Appendix

8. 1. Lexicon format

<collocate>
noun base; verbal collocate; Case of noun [;preposition] (corpus frequency)
</collocate>
<noun> noun base </noun>
<verb> verbal collocate </verb>
<lexfunc> lexical function </lexfunc>
<multilingual> French, English, and/or German translation of noun base
</multilingual>
<subcat> disjunctive list of phrase type subcategorizations with corpus frequencies
</subcat>
<noundet> disjunctive list of noun determiners with corpus frequencies </noundet>
<voice> disjunctive list of verbal voice (active, passive, etc.) of collocation
with corpus frequencies </voice>
<typical> typical (lemmatized) form of collocate </typical>
<verbalequiv> verbal equivalent of noun base
<vsubcat> phrase type subcats </vsubcat>
</verbalequiv>

8.1.2 Example entry: COMPLIMENT (pay)

<record>
<collocate> compliment ; pay; DOBJ (23) </collocate>
<noun> compliment </noun>
<verb> pay </verb>
<lexfunc> oper1 </lexfunc>
<multilingual>
<FR> compliment </FR>
<DE> Kompliment </DE>
</multilingual>
<subcat>
(7) pay compliment to
(1) pay compliment on
(1) pay compliment in
</subcat>
<noundet>
the compliment (9)
NONE compliment (7)
a compliment (3)
her compliment (2)
my compliment (1)
his compliment (1)
</noundet>

<voice>
VOICE=ACTIVE (11)
VOICE=INFINITIVE (7)
VOICE=PPART (6)
VOICE=PASSIVE (5)
</voice>
<typical>
2 ... pay compliment...
1 ... the usual compliment pay to my unimprovable English...
1 ... the ultimate compliment be pay...
1 ... the most kindly and satisfying compliment to the Cornish be pay...
1 ... the many compliment can be pay...
</typical>
<verbequiv>
compliment
<vsubcat>
(124) compliment NP
(70) compliment ... on
(20) compliment NP on
(7) compliment ... by
(6) compliment ... in
(5) compliment ... of
(5) compliment ... for
(4) compliment NP of
(4) compliment ... to
(3) compliment NP in
</vsubcat>
</verbequiv>
</record>

Notes

1 DECIDE stands for Designing and Evaluating extraction tools for Collocations In Dictionaries and corpora. This 24-month project was partly funded by DG XIII E4 of the European Commission, Luxemburg, under its MLAP-93 programme. The project (MLAP 93/19) begun in February 1994. Partners in the project were the University of Liege, the University of Stuttgart and Rank Xerox Research Centre, Grenoble.

2 The public European Commission reports mentioned in this article can be freely obtained via the WWW link, where the mutlingual lexicon produced by the project also resides: http://engdep1.philo.ulg.ac.be/decide/

3 At the University of Stuttgart, the same transformation into database form has been performed on the *Klett/Collins German-English dictionary.*

References

Abney, S. 1991. "Parsing by Chunks." In Berwick, R., Abney, S. and C. Tenny Eds) *Principle-Based Parsing*. Boston: Kluwer Academic Publishers.

Atkins, B.T.S. and A. Duval. 1978. *The Collins-Robert English-French dictionary.* Glasgow and Paris: Collins Publishers and Dictionnaries Le Robert.

Cowie, Anthony P. 1992. "Multiword Lexical Units and Communicative Language Teaching", in Arnaud Béjoint (eds), *Vocabulary and Applied Linguistics*, London: Macmillan, pp.1–12

Fontenelle, Th., Bruls, W., Thomas, L., Vanallemeersch, T. and J. Jansen. 1994a. "Comparative State-of-the-Art Survey and Assessment Study of Collocation Extraction Tools.", ms. (Liège: Université de Liège), Deliverable D-1a of the DECIDE MLAP 93/19 project. 123 pp.

Fontenelle, Th., Jansen, J. and U. Heid. 1994b. "Design and structure of the prototype collocation lexicon", ms. (Liège: Université de Liège), Deliverable D-2a of the DECIDE MLAP 93/19 project. 51 pp.

Fontenelle, Th.. October 1995. *Turning a bilingual dictionary into a lexical-semantic database*. PhD thesis. Liège: Université de Liège.

Fontenelle, Th., Alexandre, L., Gerardy, Cl., Thomas, L., and T. Vanallemeersch. 1995. "Prototype extraction tools for dictionaries", ms. (Liège: Université de Liège), Deliverable D-3b of the DECIDE MLAP 93/19 project. 48 pp.

Gerardy, C., Grefenstette, G., Heid, U., Schulze, B.M., Fontenelle, Th., Alexandre, L., and B.T.S. Atkins. 1996. "Multilingual Lexicon combining information extracted from corpora and dictionaries", ms. (Liège: Université de Liège), Deliverables D-6, 7, and 8 of the DECIDE MLAP 93/19 project. 125 pp.

Grefenstette, G.. 1994. *Explorations in Automatic Thesaurus Discovery.* Boston: Kluwer Academic Press.

Grefenstette, G., Teufel, S., Gaschler, J. and B.M. Schulze. 1994. "Specifications for collocation extraction tools", ms. (Liège: Université de Liège), Deliverable D-2b of the DECIDE MLAP 93/19 project. 42 pp.

Grefenstette, G., Schulze, B.M., Heid, U., and Th. Fontenelle. 1995. "Prototype tools for extracting collocations from corpora", ms. (Liège: Université de Liège), Deliverable D-3a I and II of the DECIDE MLAP 93/19 project. 34 and 44 pp.

Grefenstette, G. 1996. "Applying the SEXTANT parser to the IPSM Corpus." In Sutcliffe, R. (Ed.) *Low-Level Parsing applied to Technical Manuals. IPSM95,* to appear.

Mackin, R. 1978. "On Collocations: Words shall be known by the company they keep." In *In Honour of A. S. Hornby.* Oxford: Oxford University Press, pp. 149–165.

Mel'cuk, I. 1984–1988. *Dictionnaire explicatif et combinatoire du francais. Recherches lexico-semantiques.* Montreal: Les Presses de l'Université de Montreal. Vols I, II, III.

Schulze, B.M., and O. Christ. 1994. *The CQP User's Manual.* Stuttgart: Institute of Natural Languages, University of Stuttgart.

Schulze, B.M., Heid, U., Schmid, H., Schiller, A., Rooth, M., Grefenstette, G., Gaschler, J. and S. Teufel. 1994a. "Comparative State-of-the-Art Survey and Assessment Study of General Interest Corpus-Oriented Tools", ms. (Liège: Université de Liège), Deliverable D-1b I of the DECIDE MLAP 93/19 project. 95 pp.

Schulze, B.M. and U. Heid. 1994b. "State-of-the-Art Survey of Corpus Query Tools", ms. (Liège: Université de Liège), Deliverable D-1b II of the DECIDE MLAP 93/19project. 133 pp.

Wierzbicka A. 1987. *English Speech Act Verbs: A Semantic Dictionary.* Sidney: Academic Press.