Thierry FONTENELLE, Commission of the European Communities, Translation Service
Development of multilingual tools - Luxembourg

The semantic analysis of *of*-phrases for word sense disambiguation

Abstract

The aim of this paper is to show that the construction of semantic resources is a sine qua non if one wishes to tackle such complex and ambitious tasks as word sense disambiguation or (machine) translation selection, which are notorious stumbling blocks in most natural language processing systems. More specifically, I will illustrate my contention with examples featuring *of*-phrases to show that semantically-oriented lexicons are absolutely essential if one wishes to develop systems able to parse phrases and recognize their heads, a necessary step in word sense disambiguation.

Keywords: word sense disambiguation, translation selection, computational lexicography, collocation, lexical function

1. Introduction

Grefenstette (1996) shows how computers can be used to cluster a corpus of citations for a given word. Dividing citations into clusters and deciding why the cluster members belong together is a pre-requisite before the lexicographer is able to code his or her conclusions into a dictionary definition. Grefenstette argues that abstracting away surface differences in the original citations is a necessary process which can be seen as a series of successively more informed linguistic approximations to full parsing. He shows that different levels of surface abstraction, e.g. to discover the common clusters of arguments for a given verb, can be approximated by available text processing tools, ranging from simpler tools to more advanced ones:

(a) simple tokenizers, to find word boundaries;
(b) morphological analyzers/lemmatizers, to reduce inflected forms to normalized lemmata;
(c) part-of-speech taggers, to move from individual words to parts of speech, i.e. grammatical classes;
(d) low-level parsers, to discover syntactic functions, viz. direct/indirect objects, subjects, complements...

A fifth category of tools, viz. semantic taggers, is only granted a passing remark because the relative unavailability of semantic dictionaries is responsible for the relative failure of research in this area. Moreover, Grefenstette adds, the problem is no longer one of structure, but of meaning. I wish to show here that the construction of such semantic resources is a sine qua non if one wishes to tackle such complex and ambitious tasks as word sense disambiguation or (machine) translation selection, which are notorious stumbling blocks in most natural language processing systems. More specifically, I will illustrate my contention with examples featuring *of*-phrases to show that semantically-oriented lexicons are absolutely essential if one wishes to develop systems able to parse phrases and recognize their heads, a necessary step in word sense disambiguation.

2. Introduction: Lexical knowledge for natural language processing

It is a well-known fact that NLP systems such as information retrieval or machine(-aided) translation systems require knowledge about words. In order to feed the lexical components of these systems with the descriptions of tens of thousands of lexical items, researchers have concentrated on two main sources of data to overcome the so-called 'lexical acquisition bottleneck' (see amongst others, Boguraev & Briscoe 1989; Wilks et al. 1996):

(a) existing dictionaries, whether bilingual or monolingual, available in machine-readable form;
(b) large electronic textual corpora.

The main bulk of activities in this area so far has focused on the acquisition of syntactic information (subcategorization, complementation, etc.). Learners' dictionaries, whose more or less formalized grammar codes have been found so useful in this perspective, have been resorted to and analyzed extensively (see Wilks et al. 1996). More recently, attention has been paid to the extraction of co-occurrence (collocational) knowledge from corpora and from dictionaries (Church & Hanks 1990, Church et al. 1994, Fontenelle 1997a,b, Grefenstette et al. 1996, Heid 1994), in keeping with the Firthian contention that 'words shall be known by the company they keep'. Statistical methods such as mutual information (MI) calculation are now widely used by lexicographers to discover and encode in their lexical descriptions which words are more likely than others to occur in the neighbourhood of a given item (see Church et al. 1994, Clear 1993, Baugh et al. 1996). These sophisticated statistical techniques open up new perspectives insofar as they enable lexicographers and linguists to enrich their lexicons with information about the combinatorial properties of lexical items. However, these extraction programs capitalize on the recurrence of structural properties in corpora and the lists of collocates ranked in decreasing order of MI scores are usually heterogeneous sets of items. Current systems usually fail to differentiate between collocations (although Grefenstette (1994) suggests a method for doing just that in specialized lexicons) and the lexical-semantic relationship holding between a given node and its potential collocates is most often not identified or made explicit. The entry for *rose* in CIDE (Procter 1995) is a case in point. This dictionary, which used state-of-the-art techniques to tap corpus data, indicates that *rose* seems to collocate typically with *garden, bush, petal* or *bunch*. We may suppose that the CIDE lexicographer decided to include these words in the illustrative examples because s/he had found them among the statistically most significant collocates. Nothing in the MI score tables or in the dictionary entry tells us that the relationships between *rose* and the four items above are different, however. While *rose bushes* are bushes made of roses, it is crucial to realize that *petals* are part of a rose whereas *bunch* expresses a 'group' relationship in syntagmatic combination with *rose*. Making this relationship explicit would add a semantic dimension to a lexicon, which could dramatically improve the performance of a machine translation system or a reading comprehension tool as illustrated in the following section.

3. *Of*-phrases: quantifier vs. possessive phrases in a collocational perspective

The need for explicit and formalized lexical-semantic relations in dictionary entries is crucial if one wishes to design a computerized tool for providing readers of a text in a foreign language, say English, with the best possible translation of a given word, according to its environment in the source text (as in the DEFI[1] project on word sense disambiguation through filtering -

Michiels & Dufour 1996: 1). In the case of *of*-genitives, for instance, it is necessary to distinguish noun phrases such as *the roof of the house* and *a bar of chocolate* and to treat them differently.

A default approach would indeed process *N1 of N2* sequences (and the English equivalent *N2's N1*) and consider *N1* as the head of the NP. This heuristic makes it possible to treat the majority of cases in a satisfactory way and enables the system to keep track of the collocational relationship holding between this head and the predicate with which it is associated. The following sentences illustrate cases where it is indeed essential to consider *N1* as the head of the NP (underlined in the examples below):

They published the complete plays of Federico Garcia Lorca.
His collection of essays has won a prize.
The employees of the company decided to go on strike.
Few people are able to speak the three official languages of this country.
This forces us to reconsider the traditional role of women.
She clutched the sleeve of his robe.
The teacher drew a map of Sweden on the blackboard.

If an on-line reading comprehension tool (or a machine translation system) is to display the appropriate translation of a word in a given context, it should be able to establish a collocational relationship between pairs of items such as *publish* and *play*, *employee* and *go on strike*, *speak* and *language*, *reconsider* and *role*, *clutch* and *sleeve* or *draw* and *map*. Drawing on the heuristic described above, such a system would be able to do just that. In a large number of cases, however, the wrong identification of the semantic head of the NP would lead to a failure to recognize collocations which can be crucial in a word sense disambiguation perspective. This is especially true in the case of standard quantifiers (scores, hundreds, thousands, millions... of), as in the following examples (the semantic head is underlined):

These hens have laid dozens of eggs since we bought them.
Scores/Hundreds of students were examined.

Finding the correct sense, and hence the translation, of *laid* obviously depends on the system's ability to recognize a collocational relationship between *lay* and *egg*, and not between *lay* and *dozen*. At first glance, one might think that quantifiers such as *dozen, hundred, billion*, etc. belong to a finite set (including *piece*, as in *a piece of information*), and that listing a few tens of quantifiers would be sufficient to cater for most of the exceptions. The following examples, however, provide ample evidence that *N2*, the semantic head of the *of*-NP, may collocate with highly specific partitives or quantifiers:

They shot clouds/rains of arrows at us.
He offered her a bunch of roses.
Melt a bar of chocolate in hot milk.
He ate the whole tablet of chocolate.
He ate two tablets of chocolate.
I have been unable to install this piece of software.
Mary suffered from a bout of malaria / an outbreak of fever.
She was always trying to sweep up every speck/mote of dust in her house.
The teacher first wrote his name with a stick of chalk.

He wrote a collection of essays.

For disambiguation purposes, it is essential that an NLP system should be able to associate *sweep up* and *dust, offer* and *rose, shoot* and *arrow, melt/eat* and *chocolate* or *install* and *software*. It should be realized that the noun to the left of *of* is not necessarily a partitive or a quantifier in any *N1 of N2* combination, however. In the following examples, *stick, piece* and *outbreak* are the heads of their respective NPs:

The white stick of the blind woman had been broken in the accident.
Two pieces of the jigsaw puzzle were missing.
At the outbreak of the Second World War...

It should also be noted that the correct identification of the head makes it possible to keep track of adjective-noun collocational relations, as in:

a warm round of applause
a good stroke of luck
a sound piece of advice

where *warm, good* and *sound* can only be translated with reference to *applause, luck* and *advice* respectively (note that *a piece of sound advice* is possible). Michiels and Dufour (1996: 10) are right to note that this type of distinction "requires analysis of the sentence at the semantic level, which to a large extent remains the privilege of human readers". Word sense disambiguation and translation assignment is basically a semantic process, however, and my conviction was that the performance of such a system could be enhanced by resorting to an existing lexical-semantic resource such as the Liège Collins-Robert (CR) lexical-semantic database which, to some extent at least, contains the very semantic information needed in this perspective (Fontenelle 1997a,b; Atkins & Duval 1978).[2]

4. A lexical-semantic database derived from the Collins-Robert dictionary

The CR dictionary partly owes its reputation to the extensive use it makes of italicized metalinguistic information about the semantic, syntactic and combinatory properties of words. The systematic approach adopted by the CR lexicographers enables them to account for a whole range of collocational constraints and restrictions. The following examples illustrate noun-noun collocations in which the noun that usually complements a noun headword appears in square brackets:

bar 1 *n* **a** *(slab) [metal]* barre *f; [wood]* planche *f; [gold]* lingot *m*; *[chocolate]* tablette *f*
b *(rod) [window, cage]* barreau *m*; *[grate]* barre *f*, *[door]* barre, bâcle *f*, *(Sport) [ski-lift]* perche *f*
crow 1 *n [cock]* chant *m* du coq, cocorico *m*; *[baby]* gazouillis *m*; *(fig)* cri *m* de triomphe
den *n* **a** *[lion, tiger]* tanière *f*, antre *f*, *[thieves]* repaire *m*, antre.
mote *n* atome *m*; *[dust]* grain *m*
rain 1 b *(fig) [arrows, blows, bullets]* pluie *f*
sleeve 1 *n [garment]* manche *f*, *[record]* pochette *f*, *[cylinder etc]* chemise *f*
speck 1 *n [dust, soot]* grain *m*; *[dirt, mud, ink]* toute petite tache; *(on fruit, leaves, skin)* tache, tavelure *f*, *(tiny amount) [sugar, butter]* tout petit peu; *[truth etc]* grain, atome *m*

stick 1 *n* **b** *[chalk, charcoal, sealing wax, candy]* bâton *m*, morceau *m*; *[dynamite]* bâton; *[chewing gum]* tablette *f*, palette *f (Can)*; *[celery]* branche *f*; *[rhubarb]* tige *f*

The examples above illustrate various types of lexical-semantic relations, ranging from partitive relations (*a stick of candy, a stick of rhubarb...*) to quantifying (group) relations *(a rain of bullets/arrows...*) or part-whole relations (*the bars of a cage, the sleeves of a garment...*). Other relations are also illustrated, expressing typical sounds or locations (*the crow of a cock, the lion's den...*). The main problem is that the user is left in the lurch when it comes to working out the exact nature of the link which unites the various components of the collocations. In the Collins-Robert database described in Fontenelle (1997a), the semantic interpretation in question has been carried out in order to allow flexible queries and semantically-motivated questions. The relationship between the base of the collocation (the metalinguistic indicator in italics) and the collocator (which corresponds to the headword in the dictionary) has been made explicit in terms of lexical-semantic labels based on Mel'chuk's lexical functions (Mel'chuk et al. 1984 - see Fontenelle 1997a for further details on the construction of this lexical-semantic database).

The concept of lexical function is used to account for a whole range of syntagmatic (collocational) and paradigmatic relations. The notation *f(X)=Y* is used to indicate that a lexical-semantic relation *f* holds between a keyword *X* and a value *Y*. Mel'chuk's contention is that most of the systematic and recurrent lexical-semantic relationships in a general-language lexicon can be formalized in terms of a set of around 60 lexical functions. In the Liège Collins-Robert database, some of the relations illustrated above are represented as follows:

Sing (chalk) = stick
Sing (rhubarb) = stick
Sing (dust) = mote, speck
S_0Son (cock) = crow
Part (garment) = sleeve

Part (cage) = bar
Sing (gold) = bar
S_{loc} (tiger) = den
S_{loc} (lion) = den
Mult (arrow) = rain

The linguistic decisions concerning the choice of LFs have been impossible to automate and the semantic interpretation has mostly been carried out manually for over 70,000 pairs of items. This approach now makes it possible to group items which share a common meaning component. The GROUP relationship, for instance, is represented by the **Mult** LF (<Multitude), while typical nouns of sound are expressed via the **S_0Son** LF (**S_0** refers to the substantive; **Son** is used for typical verbs of sound, as in Son (dog)=bark). **S_{loc}** stands for substantives denoting typical locations and **Sing** expresses regular portions or units of something (*a piece of information, a grain of rice, a blade of grass...*). As can be seen, **Mult** and **Sing** are of special interest here because they account for the very relations we seek to identify. *A speck/mote of dust, a bunch of roses, a rain of arrows, a bar/tablet of chocolate* ... are all combinations which can be expressed in terms of these two LFs and for which the values (Y in the mathematical notation described above) are specific instances of quantifiers. What makes this approach interesting in a word sense disambiguation perspective is that the LFs hold between pairs of items, which means that different functions may be used for a given dictionary headword, depending on the metalinguistic indicators with which it is combined. Consider *bar*, for instance, which is the exponent of the **Part** function when associated with *cage, door* or *window*, but is the value of the **Sing** function when associated with *chocolate, gold*, etc.

Coding all these relations across the entire dictionary now makes it possible to answer queries such as: List nouns which express a regular portion/unit of 'chocolate', which boils down to searching the database for nouns linked to an italicized reference to *chocolate* by means of the **Sing** lexical function. The CR database contains the following information:

Sing (chocolate) = bar, cake, piece, tablet

Even if the database does not make any distinction between *a bar/tablet of chocolate* and *a piece of chocolate* (the former combinations refer to standard, marketed units which are themselves divisible into smaller pieces), it must be stressed that, in the perspective adopted here, this information is amply sufficient to consider *chocolate* as a head in these combinations and the various collocators as quantifiers. Since a large lexical resource was available, the DEFI team was able to use the lexical functions **Sing** and **Mult** as filters, bearing in mind that the database obviously does not and cannot include all possible combinations featuring quantifiers (consider the use of *etc* in **speck** *n [truth etc]* grain, atome), but that the lists which can be retrieved comprise over 1,000 such pairs, which by far exceeds what can be found in standard grammars. The lists in Table 1 give an idea of what can be found in the CR database.

Table 1. Sing and Mult in the Collins-Robert database

Sing (sample list)	Mult (sample list)
sing (air / air) = puff (bouffée <f>)	mult (abuse / injure) = spate (torrent <m>)
sing (air / air) = puff (souffle)	mult (abuse / injure) = storm (torrent <m>)
sing (air / air) = waft ((petite) bouffée <f>)	mult (ant / fourmi) = nest (nichée <f>)
sing (air / air) = whiff (bouffée <f>)	mult (ant / fourmi) = swarm (fourmillement <m>)
sing (alcohol / alcool) = drop (goutte)	mult (applause / applaudissement) = burst (salve <f>)
sing (alcohol / alcool) = shot (coup <m>)	mult (applause / applaudissement) = storm (tempête)
sing (amber / ambre) = bead (perle <f>)	mult (applause / applaudissement) = thunder(tonnerre)
sing (amber / ambre) = bead (grain <m>)	mult (applause / applaudissement) = volley (salve)
sing (anger / colère) = access (accès <m>)	mult (arrow / flèche) = cloud (nuée)
sing (anger / colère) = burst (explosion)	mult (arrow / flèche) = rain (pluie <f>)
sing (anger / colère) = burst (éclat <m>)	mult (arrow / flèche) = sheaf (faisceau <m>)
sing (anger / colère) = eruption (explosion <f>)	mult (arrow / flèche) = shower (pluie <f>)
sing (anger / colère) = outbreak (explosion <f>)	mult (arrow / flèche) = storm (pluie <f>)
sing (anger / colère) = outburst (explosion)	mult (asparagus / asperge) = bunch (botte <f>)
sing (beef / boeuf) = quarter (quartier)	mult (banana / banane) = bunch (régime <m>)
sing (beer / bière) = pint (demi <m> (de bière))	mult (banana / banane) = cluster (régime <m>)
sing (billiards / billard) = game (partie)	mult (banknote / billet de banque) = roll (liasse <f>)
sing (bread / pain) = bit (morceau <m>)	mult (banknote / billet de banque) = wad (liasse)
sing (bread / pain) = chunk (quignon <m>)	mult (bee / abeille) = cluster (essaim <m>)
sing (bread / pain) = pellet (boulette <f>)	mult (bee / abeille) = swarm (essaim <m>)
sing (bread / pain) = piece (morceau)	mult (bullet / balle) = rain (pluie <f>)
sing (bread / pain) = round (tranche <f>)	mult (camel / chameau) = train (caravane <f>)
sing (bread / pain) = scrap ((petit) bout <m>)	mult (cattle / bétail) = herd (troupeau <m>)
sing (bread / pain) = slice (tranche <f>)	mult (cheer / applaudissement) = storm (tempête <f>)
sing (cheese / fromage) = knob (petit morceau)	mult (chick / poussin) = hatching (couvée <f>)
sing (cheese / fromage) = lump (morceau)	mult (chicken / poulet) = clutch (couvée <f>)
sing (cheese / fromage) = sliver (lamelle <f>)	mult (conscript / conscrit) = draft (contingent <m>)
sing (chess / échecs) = game (partie)	mult (corn / blé) = sheaf (gerbe <f>)
sing (dew / rosée) = bead (perle <f>)	mult (curse / juron) = stream (flot)
sing (malaria / malaria) = bout (attaque)	mult (curse / juron) = string (kyrielle <f>)

The availability of this database made it possible to reuse the semantic triples it contains in order to improve the performance of the genitive-analysis routine used in the DEFI project. The algorithm used by the DEFI analyzer for identifying the basic relation between the two nouns in a genitive NP and for computing the semantic head is very simple. The routine takes two arguments corresponding to *N1* and *N2*. To compute the relation between *N1* and *N2*, it first checks whether the combination is attested in the CR database. In other words, it searches for occurrences of *N1* in the field containing headwords and for occurrences of *N2* in the field containing metalinguistic indicators. *N1 of N2* structures are regarded as genitive phrases only if no **Mult** or **Sing** relationship is found between *N1* and *N2* (see Michiels & Dufour 1996). Consider the following entry from the printed dictionary:

blade *n [knife, tool, weapon, razor]* lame *f; [chopper, guillotine]* couperet *m; [tongue]* dos *m; [oar]* plat *m*, pale *f; [spade]* fer *m; [turbine motor]* aube *f; [propeller]* pale, aile *f; [windscreen wiper]* caoutchouc *m*, balai *m; [grass, mace]* brin *m; [cereal]* pousse *f; [leaf]* limbe *f*

Consider the following *of*-phrases to be interpreted:

1. a blade of grass
2. the blade of a razor

When the system finds a reference to *N2* under *N1*, it checks whether a lexical function has been assigned to account for the lexical-semantic relation between *N1* and *N2*. If either **Sing** or **Mult** has been assigned, which is the case for *blade [of] grass* above (Sing(grass)=blade), the routine considers *N1* as a quantifier and *N2* as the head of the NP (*blade* = quantifier ⇒ head = *grass*).

If, however, another function is assigned, or no function at all, the system considers *N1* as the head of the genitive NP (*blade of a razor* ⇒ head = *blade*). The decision here is based upon the presence of a function **Part** indicating that a part-whole relation holds between *blade* and *razor*.

In order to cope with standard quantifiers which can be combined with any noun whatsoever and which therefore do not enter lexically restricted collocations (*dozens of people, billions of dollars, hundreds of books, 250 grams of cheese...*), ad-hoc solutions have been foreseen to force the system to consider *N2* as the semantic head when *N1* belongs to a set including *dozen, hundred, thousand, million* and a few others (standard units such as *kilo, gram, litre...*).

It should of course be acknowledged that this algorithm will not be able to identify the heads of an *of*-phrase with 100% reliability. In some cases, certain phrases should indeed be analysed differently, depending on the context. In *He drank a whole bottle of whisky*, *whisky* may be considered as the semantic head and enters into a collocational (or at least semantic) relationship with the verb *drink*. In *He dropped the bottle of whisky, which broke*, the noun *bottle* is the semantic head of the *of*-phrase and the verbs *drop* and *break* should be disambiguated with respect to *bottle*. Such cases of ambiguities fall within the scope of Pustejovsky's Lexical Conceptual Paradigms (e.g. the container/content alternation) and should probably be solved on the basis of his so-called qualia structures (Pustejovsky 1991).

5. Expanding the lists

All the above examples make it abundantly clear that collocations do not go hand in hand with syntactic structures. In the sentence *His collection of essays has won a prize*, the quantifier *collection* controls agreement. Obviously, other syntactic clues can also be used in order to help the disambiguation procedure. Most of the cases illustrating a quantifier *of*-phrase indeed feature the following structure:

(the/a/an/∅) N1 of N2

where N2 is either a singular uncountable noun or a plural countable noun. Most of the time, N2 is not preceded by any determiner (*a bout of malaria*), which means that the presence of a determiner between *of* and N2 can be reasonably considered as a clue signalling a possessive phrase (*the employees of the company*). Such clues were used in order to exploit the examples of the dictionary, which had not been used in the original lexical-semantic database. The purpose was to expand the lists by discovering cases of quantifier phrases nested in illustrative examples. Consider the following example excerpted from the printed dictionary:

bout *n (period)* période *f; [malaria etc]* attaque *f,* accès *m*; **bout of rheumatism** crise de rhumatisme; **bout of fever** accès de fièvre; **a bout of bronchitis**; une bronchite; **a bout of flu** une grippe...

As explained above, only italicized collocations were taken into account when constructing the lexical-database and enriching it with lexical functions. This means that the LF Sing only linked *bout* and *malaria*. There is no principled reason to exclude the other bases which are hidden in the illustrative examples, however. All the dictionary examples containing *of* were therefore extracted, with a view to retrieving additional pairs of Sing/Mult collocations. The small sample in Table 2 shows that the resulting list contains noise and that it is necessary to separate the wheat (to the left) from the chaff (to the right).

Table 2. *Of*-phrases in dictionary examples (sample list)

a bout of bronchitis (une bronchite)	**a letter of apology** (une lettre d'excuses)
a bout of flu (une grippe)	**a nod of approval** (un signe de tête approbateur)
a piece of china (une porcelaine)	**balance of payments** (balance des paiements)
a crumb of comfort (un brin de réconfort)	**bill of sale** (acte/contrat de vente)
bunch of flowers (bouquet (de fleurs))	**betrayal of trust** (abus de confiance)
bunch of grapes (grappe de raisins)	**look of astonishment** (regard stupéfait)
burst of rain (averse)	**a breach of decorum** (une inconvenance)

The assignment of the function Sing or Mult to pairs of collocates in the entry corresponding to N1 was then used to filter out irrelevant examples. The hypothesis was that the Sing or Mult function used to code the relationship between N1 and other nouns granted quantifier status to N1 and that this property could be percolated down to the other nouns with which it cooccurred, provided they appeared in the dictionary under the form of a well-defined structure (a/an/the/∅ N1 of ∅ N2, where ∅ stands for the zero article). Since **bunch** is associated by means of the Mult function with *tulip, rose, asparagus, radish, banana...,* this hypothesis made it possible to extend its semantic property to other nouns mentioned in the example section such as *flowers* and *grapes* which had not been considered in the original work on the database. Conversely, the absence of any such reference to Sing or Mult in the entry for *letter, nod, balance, bark, betrayal, look* or *breach* was sufficient to exclude the combinations which appear in the left-hand column above.

6. Conclusion

The problems addressed in this paper make it abundantly clear that NLP lexicons require a semantic dimension which is all too often lacking in existing systems. The approach sketched here only solves one particular type of problem and the development of flexible and robust tools for syntactic analysis clearly requires access to syntactic and semantic information about words which goes far beyond the formalization of the two lexical relations for quantifiers and partitives. Parsers need to be able to compute the semantic heads of phrases which stand in argument relation and their performance should therefore be dramatically increased when the purely semantic information available in dictionaries is made fully explicit and formalized, as I have tried to show in this paper.

7. Notes

[1] DEFI is a nationally-funded five-year Belgian project run at the University of Liège in the field of word sense discrimination and target selection in bilingual dictionaries (1995-2000). The DEFI technical reports are available from the following URL: http://engdep1.philo.ulg.ac.be/michiels/defi.htm

[2] The Collins-Robert lexical-semantic database referred to in this paper was developed in the framework of a doctoral dissertation while the author was working in the English Department of the University of Liège. Part of the development of this copyrighted database was financed by the European Commission in the framework of the DECIDE project (MLAP 93/19). The DECIDE reports and deliverables can be freely obtained by the WWW link (http://engdep1.philo.ulg.ac.be/decide/). Examples illustrating the multiple access keys provided by the CR query program can be found at the following address: http:// engdep1.philo.ulg.ac.be/decide/robcol-examples.html.

8. References

Atkins, B.T.S. & Duval, A. (1978). *Collins-Robert English-French French-English Dictionary*, HarperCollins Publishers and Dictionnaires Le Robert, Glasgow & Paris.

Baugh, S., Harley, A., Jellis, S. (1996). The Role of Corpora in Compiling the Cambridge Dictionary of English, in *International Journal of Corpus Linguistics*, 1/1, pp.39-59.

Boguraev, B. & Briscoe, T. (1989). *Computational Lexicography for Natural Language Processing*, London and New York, Longman.

Church, K., Gale, W., Hanks, P., Hindle, D. & Moon, R. (1994). Lexical Substitutability, in Atkins & Zampolli (eds) *Computational Approaches to the Lexicon*, Oxford University Press, pp.153-177.

Church, K. & Hanks, P. (1990). Word Association Norms, Mutual Information and Lexicography, in *Computational Linguistics*, 16/3, pp.22-29.

Clear, J. (1993). From Firth Principles: Computational Tools for the Study of Collocations, in Baker, Francis & Tognini-Bonelli (eds) *Text and Technology - In Honour of John Sinclair*, Amsterdam & Philadelphia, Benjamins, pp.271-292.

Cruse, D. (1986). *Lexical Semantics*, Cambridge University Press.

ENGCG - Constraint-based grammar developed at the General Linguistics Department of the University of Helsinki. It is marketed by Lingsoft Inc. (http://www.lingsoft.fi).

Fontenelle, Th. (1997a). *Turning a bilingual dictionary into a lexical-semantic database*, Lexicographica Series Maior, Band 79, Max Niemeyer Verlag, Tübingen.

Fontenelle, Th. (1997b). Using a bilingual dictionary to create semantic networks, in *International Journal of Lexicography*, 10/4, pp.275-303.

Grefenstette, G. (1994). Corpus-Derived First, Second and Third-Order Word Affinities, in *EURALEX'94 Proceedings*, Vrije Universiteit Amsterdam, pp.279-290.

Grefenstette, G., Heid, U., Schulze, B.M., Fontenelle, Th., Gérardy, C. (1996). The DECIDE Project: Multilingual Collocation Extraction, *EURALEX'96 Proceedings*, Göteborg University, pp.93-107.

Grefenstette, G. (1996). Approximate Linguistics, in *Proceedings of the 4th Conference on Computational Lexicography and Text Research - COMPLEX'96*, Budapest, Hungary, Sept.1996.

Heid, U. (1994). On Ways Words Work Together - Topics in Lexical Combinatorics, in *EURALEX'94 Proceedings*, Vrije Universiteit Amsterdam, pp.226-257.

Mel'chuk, I. et al. (1984). *Dictionnaire explicatif et combinatoire du français contemporain*, Presses de l'Université de Montréal.

Michiels, A., Dufour, N. (1996). *From SGML tape to DIC clauses: Identifying Multi-Word Units for Context-Sensitive Lookup*. DEFI Technical Report 2, CIPL, Liège. (available from the following URL: http://engdep1.philo.ulg.ac.be/michiels/defi.htm).

Miller, G., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. (1990). Introduction to WordNet: An On-Line Lexical Database, in *International Journal of Lexicography*, 3/4, pp.235-244.

Procter, P. (ed.) (1995). *Cambridge International Dictionary of English*, Cambridge University Press.

Pustejovsky, J. (1991). The Generative Lexicon, in *Computational Linguistics*, 17/4, pp.409-441.

Wilks, Y., Slator, B., Guthrie, L. (1996). *Electric Words - Dictionaries, Computers and Meanings*, MIT Press, Cambridge, MA and London.