Collocations, Dictionaries and MT

Dirk Heylen
heylen@let.ruu.nl
OTS, Utrecht University, The Netherlands
Kerry G. Maxwell
kerry@essex.ac.uk
CL/MT Group, University of Essex, United Kingdom
Susan Armstrong-Warwick
susan@divsun.unige.ch
ISSCO, University of Geneva, Switzerland

1 Introduction

Collocations pose specific problems in translation (both human and machine translation). For the native speaker of English it may be obvious that you 'pay attention', but for a native speaker of Dutch it would have been much simpler if in English people 'donated attention.' Within an MT system, we can deal with these mismatches in different ways. Simply adding the entry to our bilingual dictionary saying 'pay' is the translation of 'schenken', leaves us with the job of specifying in which contexts we can use this *equivalence*. A more elaborate dictionary might list the complete collocation alongside with its translation. Another possibility would be to adopt an interlingua and have (harmonious) monolingual components take care of mapping 'pay' and 'schenken' to the same meaning representation. The idea of investigating the latter possibility suggested itself while looking at the analysis of collocations in terms of lexical functions by Mel'čuk and his followers (as presented for instance in their Explanatory Combinatory Dictionaries [Mel'čuk *et al.*, 1984]). The claim that these lexical functions are universal and only few in number, would make them interesting for the interlingua option. The ET-10 project 'Collocations or the lexicalisation of semantic operations'[1] is concerned with investigating these claims and their consequences for the automatic translation of collocations.

Adopting this approach, leads rather naturally towards an interlingua system, using the lexical functions as some sort of *semantic primitives*. Being rather sceptical about the use of such concepts, we set ourselves the task to find out whether the 'empirical basis' (collocations) constitute sufficient ground to postulate this specific inventory of semantic primitives. This has lead us into the investigation of corpora and dictionaries to confront abstract concepts with real data. Although not one of our tasks, these investigations have lead to a number of suggestions on how one could possibly extract relevant information from these sources for the construction of collocational dictionaries. In this paper we will focus our attention on the methods to extract information from dictionaries.

[1] The research project (ET-10/75) is sponsored by the European Commission, L'Association Suissetra and Oxford University Press.

2 Collocations and Semantics

Previous work ([Church and Hanks, 1989], [Smadja, to appear] and earlier work) on the acquisition of collocational information has concentrated on the extraction of collocations from corpora based on some statistical measures. Programs such as XTRACT ([Smadja, to appear]) are able to select an interesting collection of potential 'collocations'. The notion of collocation that lies behind these investigations is that of 'recurrent combination'. The notion we are investigating however, is different as it only concerns collocations in which we can clearly distinguish a collocate which has a 'general' or 'non-literal' reading. The mode of defining 'collocations' in our approach is different from the *p*erformance or *p*robability approach. Instead, we are interested in a *l*inguistic or *l*exicographic way of defining collocations and more particularly the semantics of collocations. The first approach leads to the notion of collocations we have termed *p-collocation* elsewhere ([Bloksma *et al.*, 1992]), whereas the second mode leads to a set of *l-collocations*[2].

But what are the semantic aspects of collocations we are interested in exactly? In the analysis by Mel'čuk of the kind of collocations we are interested in (mainly adjective-noun combinations and light verb constructions), a construction such as *heavy smoker* is analyzed by means of a lexical function **Magn**. In the Explanatory Combinatory Dictionary[3] the entry for *smoker* contains information of the following kind.

$$\mathbf{Magn}(\mathit{smoker}) = \mathit{heavy}$$

Mel'čuk claims that we need about 50 of these 'universal' functions to describe the semantics of most collocations[4]. A lexical function represents some meaningful operation which yields the lexeme (*collocate*) that typically combines with the argument (*head* of the collocation) of the function to express the meaning of the function. Simplifying, **Magn** — meaning something like "intense" — is expressed by *heavy* when said of *smoker*. More complex functions can be constructed by (1) combining functions; (2) adding different kinds of sub- and superscripts. The following examples illustrate this.

$$\mathbf{IncepFunc}_1(\mathit{vent}) = \mathit{se\ lever}$$
$$\mathbf{FinFunc}_1(\mathit{vent}) = \mathit{se\ calmer}$$
$$\mathbf{Magn}_{['consequences']}(\mathit{maladie}) = \mathit{sérieuse}$$

The last example, clearly illustrates that the basic lexical functions are in need of some refinement. The semantic component of the argument which the intensifier relates to is explicitly mentioned in the subscript. In [Heylen, 1992] we have suggested some ways to make lexical functions sensitive to the semantic structure of the arguments, by incorporating for instance *qualia* information ([Pustejovsky, 1991]) in the semantic representation of nouns. Notice also that lexical functions go with the ideology that the association between heads and collocates is to be specified lexically, because it is unpredictable. They do not consider any kind of (sub)regularities that may exist.

One of the aims of the project is to evaluate Mel'čuk's system of lexical functions. Among the questions we are interested in are the following.

- Can collocations[5] be associated with lexical functions as the theory assumes they can?
- In what ways can the notion of lexical function be refined?
  - Can we make them sensitive to the semantic structure of their arguments?
  - Can we make precise the parameters that differentiate the possible values of a function? For instance, we might want to differentiate between *sharp/wide turns*, or for instance between certain "degrees of intensity" when we are considering **Magn** etc.

[2] In fact, it turns out that a further distinction should be made between the lexicographic and the linguistic approach to collocations. We will not go into further details here.

[3] The dictionary that goes with the Meaning Text Model.

[4] Understood as l-collocations.

[5] At least a definable subset.

- Can we find some regular patterns in the formation of collocations?

From a methodological point of view, it is important to be able to go beyond possibly deceiving intuitions. We therefore have to provide for more reliable ways to get at our data. Listings of collocations as provided by XTRACT-like programs, are useful as filters on the data to consider. But as we are concerned predominantly with semantic issues when evaluating the lexical function approach, we need programs that can be made sensitive to the meaning of the expressions considered. To get at a more interesting collection of data, we have looked for material from which certain aspects of meaning can be read off in one way or another. Of course, dictionaries are one source which deals directly with the representation of meaning. But it is not the explicit encoding which is most interesting to look at. Indeed, it turns out that most dictionaries do not provide the kind of information we are interested in directly[6]. It is the less obvious structure of the dictionary entry which gives rise to interesting pieces of information. In the following sections we list a number of ways we can use dictionaries to collect data on the semantic issues we are concerned with.

3 Monolingual Dictionaries

In this section we investigate how far we can gather information about collocations and collocational behaviour by scrutinizing the formats of learner's dictionaries. We focus our attention on the Oxford Advanced Learner's Dictionary, third edition [Hornby, 1974] (henceforth OALD3), but we would expect these observations to have more general validity. We examine the dictionary on both a macro- and microlexicographical level, ie: in addition to looking at the details of entry organisation we investigate the lexicographic conventions adopted and the overall rationale behind the organization of the dictionary.

3.1 Lexicographic Conventions

Of course in producing the dictionary a set of rules and general practices have to be developed in order to ensure internal consistency. In other words, where there is perceived flexibility in description/methodology, the lexicographer may adopt a certain path simply because this corresponds to a convention among those established for the construction of the particular dictionary. Some of these conventions will ultimately reflect linguistic facts, whilst others will be purely arbitrary. Such conventions often remain implicit in the nature of the resulting dictionary and are never directly explained, whilst others may occasionally be alluded to in the frontmatter of the dictionary. For instance, in the frontmatter of OALD3 we find an explanation of the listing of definitions and meanings of words, where the user is told that:-

> "...definitions are listed in order of meanings from the most common or most simple to the most rare or most complicated" (p.xiv)

Such a convention may seem arbitrary[7] in the sense that it seems to rest on the intuitions of individual lexicographers, but what might it imply in our quest for information on the semantics of collocates? Of course we should note that "rare" and "complicated" and "simple" and "common" are not always mutually prevailing characteristics, ie: a common meaning is not necesarily simple nor a rare meaning complicated. A useful interpretation for our purposes is perhaps that "rare/complicated" corresponds to "figurative/nonstandard" and accordingly "simple/common" corresponds to

[6] First of all, we are interested in 'abstract' semantic concept. Secondly, lexicographic tradition treats a number of different constructions in the same way they treat the kind of collocations we are interested in.

[7] However we should point out that lexicographers work to a set of explicit compiler instructions which may determine the most arbitrary decisions. Further, lexicographers should be highly skilled with rich intuitions about the language under description.

"standard" or "regular", so that in the case of entries for adjectives which form collocates, the definitions which constitute the secondary, or later explanations of meaning (ie: those that occur "further down" the entry) will be the more significant ones for our interests.

In the discussion of the representation of idiom structures, we are told that:-

> "..to find an idiom, look for it in the entry for the most important word in the phrase or sentence (usually a noun, verb or adjective)." (p. xvi)

This means for instance that the idiom ***pick holes in*** can be found in the entry for ***hole*** and that the idiom ***get hold of the wrong end of the stick*** can be found under the entry for ***stick***. Of course, most "important" merely corresponds to what is perceived to be most salient, though it may be that observation of what are considered to be the most salient elements of idioms may lead us to intuitions about where the semantic load lies in such structures.

In sum, even though the conventions underlying the description of such lexical phenomena may appear arbitrary at first sight, it may be that some reveal general linguistic insights, explain omissions, or provide pointers to the information in an entry that we are most interested in. We should at least be aware of them as we embark upon extraction of information, automatic or otherwise.

3.2 User Needs

A natural consequence of the learner orientation of the OALD3 is that the dictionary weighs heavily on explanatory and exemplary data, generally residing in specific example fields, and, as will be shown in later sections, this field of information proves extremely fruitful in the collection of collocational data. The frontmatter details several reasons for the extensive inclusion of example fields (p. xvii), the key one being of course their general role in reinforcing the explanation of the meaning of (groups of) words. Also worthy of note is the aim of illustrating words and multiword forms in different sentence patterns[8], and the indication of the "style" or "context" in which the word or phrase is usually used. The latter means that example fields often:-

> "...include words or sorts of words that the headword is usually used with... for example at **sensational (2)** there is *a sensational writer/newspaper.*" (p. xviii)

Such examples illustrate the potential for stumbling across example collocations, as well as hints about lexical selection, which may in turn reveal issues in the semantics of collocational structures (cf: section 3.3.1).

To summarize - an integral aspect of the rationale behind the dictionary is its exemplary function in the learner context. This means that information of relevance to the study of collocations may be implicit in exemplary form, as well as in standard headword and definition formats.

3.3 Collocational Information within Entries

In this section we will investigate to what extent information about collocational structures is available in OALD3 entries. The observations made in this section are based on a study of the electronic version of the dictionary (henceforth OALD3E) undertaken during our project. For further documentation of the results see [Heylen *et al.*, 1992]. The electronic version of the dictionary contains entries whose information fields are tagged by

[8] This information may be patchy where the facts about collocational structures are concerned; for instance in the entry for *attention* OALD3 illustrates that the related support verb construction can be passivised,eg: *No attention was paid to my advice*, but says nothing in the entry for *heavy* about restrictions on predicative use where *heavy* has a collocational reading ie: *this smoker is heavy* can only mean that the smoker has weight problems! Such gaps in coverage often arise where dictionaries concentrate on providing positive rather than negative information, ie: providing well-formed examples but not giving explicit guidance as to what is not possible.

SGML-style markers. A simple search tool developed in-house was used to extract entries bearing a headword or example matching the specified search string[9].

The focus for this study was an investigation of the coverage of Adjective Noun (henceforth AN) collocations in OALD3E. In particular we investigate the information available in adjective entries, these proving to be a more fruitful source of interesting AN structures than noun entries. Adjective entries tend to list typical (nominal) modificands, the reverse is not necessarily true (presumably since the set of modifiers given a particular noun is potentially much larger and heterogenous). Further, adjective entries provide us with secondary sense definitions and corresponding examples which may relate directly to semantic aspects of collocational behaviour, as we discuss below.

3.3.1 Searches on Adjective

The structures/combinations we are primarily interested in are those which can be deemed to be collocational on some kind of linguistic basis. For AN-collocations we refer to the noun 'base', a semantically autonomous element preserving its standard interpretation, and an adjectival 'collocate', an adjective which has some kind of figurative or non-standard reading, its precise interpretation being only derivable in the context of the noun base. These being the kind of phenomena we are attempting to investigate, it would seem that we need to be particularly attentive to those aspects of an adjective entry devoted to explanations of secondary (or further alternative) senses. Such explanations give clues as to the potential collocational readings of an adjective, and related example fields contain candidate collocational structures. An example would be the OALD3 description of primary and alternative senses for the adjective *violent*, and related examples, as shown below:-

1. using, showing, accompanied by, great force
violent wind/attack/blows/temper/abuse
2. caused by violent attack
violent death
3. severe
violent toothache

The example details three senses of the adjective. We might designate collocational status to the third, *violent* assumes an intensifying function not explicitly related to its regular use indicating `force` (cf: sense **1**).

Clues about collocational readings are also found in glosses to individual examples, eg: we see the gloss `extreme` against the sense of *violent* in the example combination *violent contrast*, where we might assume the adjective has a similar function to that in sense **3** above.

So, although there is no explicit discrimination of "collocational readings" listed in the dictionary, in sifting through the various senses of adjectives we find a number of candidate collocates (and indeed example collocations).

But it may not be merely the semantics of the adjectival collocate which is alluded to in glosses and secondary senses. We might assume that the semantics of a noun is crucially related to such factors as form, function and causation (cf: the notion of *Qualia* as coined by Pustejovsky in a discussion of lexical semantic decomposition of nouns [Pustejovsky, 1991]). These meaning aspects may also be reflected in sense definitions pertaining to particular examples where the noun appears. For instance, the sense definition against the example *violent death* says "caused by violent attack", reflecting the fact that death essentially involves a "cause" of some kind (cf: Pustejovsky's *Agentive* role). The third sense of the adjective *regular* (cf: below) refers to explicit properties of the nouns provided in the example, eg: "qualified" and "trained" of *soldiers/army*, (cf: Pustejovsky's *formal* role).

3. properly qualified; recognised; trained
regular soldiers/army

Example fields often group potential modificands of a specified adjective according

[9] Based on a Perl regular expression, cf: [Wall and Schwartz, 1990], pp.103-106.

to some unwritten semantic criteria (which sometimes may be linked back to a particular sense definition). For instance the secondary definition of the adjective *formidable* is supplied as "requiring great effort to deal with or overcome", the modificands *obstacle/opposition/enemies/debts* are supplied on mass, each in some sense representing a situation or body which must be confronted and ultimately surmounted or "overcome". So we observe that example fields frequently list sets of modificands which may be perceived as belonging to a homogenous semantic class, ie: they have a common property, or, more formally, they share a semantic feature. We may speculate about what this feature is, but, as we illustrated, the sense definition of the adjective may give clues. In some sense then we gain pointers as to potential selectional restrictions between an adjectival modifier and the set of exemplary modificands. Certain examples seem to show adjectives predicating of nouns which share particular properties.

Beyond sense definitions and related examples we also find candidate collocations in information fields which are intended to house complex expressions such as idioms or compound forms. An example is the entry for the adjective *blind*, where we find the examples *blind spot, blind turning, blind flying, blind alley, blind date.* These complex expressions may undoubtedly warrant various labels according to a given classification, but it may be that some fit our criteria for collocational status. Eg: if we take *blind flying* and *blind turning*, we observe that both have a base with its regular compositional meaning, and a modifier which derives its precise interpretation in the context of the base[10]. Such examples indicate that we should also target compound group fields as a potential "quarry" of candidate collocations.

If a collocational form occurs within an information slot devoted to compounds or fixed expressions, this may give clues as to the degree of syntactic/morphological versatility displayed by the collocate. In other words, there are certain AN combinations which we may classify as collocational according to our basic perception, but which resemble compounds/fixed expressions in terms of eg: the lack of morphological versatility (comparative/superlative formation) of the non-head. Such examples typically reside in compound group fields (eg: *tall/*taller/*tallest story, tall story* being listed in the compound group field of the OALD3 entry for the adjective *tall*). By contrast, those collocations higher on the versatility[11] scale seem more likely to crop up in example fields eg: *heavy rain.*

3.4 Summary and Conclusions

In sum, the observations made in the previous sections show that information on collocational structures can be obtained from an on-line dictionary, but that this information is practically always implicit, embedded in aspects of the formal structure of the dictionary. We have performed automatic searches for particular strings and then hand-filtered information from the entries which formed the output. The result of such hand-filtering reveals where the "hot spots" of information lie.

Various factors interact to place information at a certain place and present information in a certain way. The information gleaned from such 'first-pass' searches, coupled with a knowledge of underlying lexicographic conventions and the rationale behind the dictionary (ie: user needs, look-up strategies), can promote the simulation of more "intelligent" searches at a subsequent stage. For instance, we now know that example fields will probably form one of the most productive active search spaces. If we want data on adjective-noun collocations, we should target adjective entries primarily. Example collocations are likely to be found at example slots for sense definitions

[10] *blind* in *blind turning* means "not easily seen by drivers" and *blind* in *blind flying* means "unable to see due to cloud/fog, therefore flying with the aid of instruments only".

[11] By 'versatile' we may also refer to syntactic aspects, eg: predicative use: *the rain was heavy/*the story was tall*

which are "lower down" the entry. It is in these later sense definitions that we may find the description of senses of adjectival collocates, from which we may be able to unpack some kind of semantic information. If example modificands are 'lumped' together, (commonly delimited by /) then we may have turned up a potential selectional restriction between modifier and head (collocate and base)... and so on. We therefore illustrate how on-line lexical information, though only partially formal in nature, can be the focus of automatic searches which are crucially based on some kind of "intelligent" strategy derived from knowledge of the nature of the dictionary.

4 Bilingual Dictionaries

Examination of bilingual dictionaries can be telling in various ways, not only in the collection of contrastive data but also in the affirmation of ideas about the nature of collocational behaviours. Of course contrastive analysis of collocational structures reveals a degree of complexity in translation, but it can also bring to light regularities about the monolingual system. In addition the data can help instantiate or repudiate definitions of interlingua concepts in view of translation.

In this section we will discuss what types of information we can find in bilingual dictionaries and how this might help us towards a more systematic account of collocations (in view of automatic translation).

Though the investigation of collocations has not been a focal point for past work with machine readable dictionaries, we can identify a number of related studies which have focussed on semantics and translation issues. Bilingual studies carried out in the ACQUILEX project, for example, have investigated the linking of semantic hierarchies which have been built on language internal concepts [Sanfilippo *et al.*, 1992]. In a comparaison of noun taxonomies cross-linguistically (using word senses and hierarchies derived from mono- and bilingual dictionaries) [Vossen, 1991] defends the hypothesis that "meaning is a language internal affair". Translation equivalence is seen as a mapping between different conceptualizations. Though the work does not focus on collocations, it does recognize that a major problem remains in how to select the best language expression for a given conceptualization. It is precisely this problem that we are trying to account for with the lexical functions.

Other studies of interest in view of extracting information from bilingual machine readable dictionaries includes the work reported on in [Calzolari, 1983] on establishing semantic links and [Byrd *et al.*, 1987] on mapping between entries and problems of symmetry. The monolingual dictionary work on navigating through entries reported on in [Chodorow and Byrd, 1985] provides the basis for our work on exploring the data in bilingual dictionaries.

4.1 Accessing the data

The bilingual dictionaries we have available on-line are the Collins German, French and English, pocket edition [Collins, 1990] in all respective pairs. The dictionaries cover the basic vocabulary of the languages (ca. 15,000 words per direction) and include the major sense distinctions (differentiated only in so far as necessary for translation purposes). Though relatively small, the advantages are that they are easy to parse and contain only minimal but essential information. Thus, for a first attempt at organizing some of the data in view of our goals to systematically account for the relations between the words in collocational phrases, they provide an ideal starting point. As suggested above in the discussion of monolingual dictionaries, we assume that the data contained in these dictionaries reflects an initial synthesis and organization of essential semantic and translational considerations regarding the use of the words.

The dictionaries under consideration can be accessed via a network based dictionary consultation tool developed at ISSCO for the University of Geneva [Petitpierre and Robert, 1991]. The underlying format is an SGML[12] mark-up of the fields, identifying headwords, grammatical information, sense distinctions, contextual information, translations, example phrases, etc. This structure allows access to the entries not only by headword, but also by words found in the different fields (via secondary indexes).

4.2 Collocations and translation

As was the case with the monolingual information found in different fields, glosses and sense definitions in bilingual entries play a crucial role in providing more information about the semantics of the collocate. From the user perspective, they are of course there to assist the choice of the correct translation, ie: they are linked to the specification of a particular translation where several alternatives are possible. From our perspective, they give pointers to the possible interpretation of the collocate. Eg: in the Collins French-English/English-French Dictionary [Collins, 1990] we see *(gain, require, bring/carry, etc)* listed as senses of the verb *take*. It is precisely these sense definitions which may give clues as to the interlingual concepts which underlie such support verbs, concepts which are independent of their realisation in a given language. Eg. *take* in English translated as *remporter* in French, but meaning GAIN in both.

4.2.1 Bilingual lexicographic conventions

There are a number of conventions (often implicit) used in bilingual lexicographical work which come under the global heading of "semantic indicators". The intention is to identify or restrict the semantic range of the word to be translated. Aside from the closed class of subject field markers (eg. NAUT, GEO, MED, etc.), these *semantic indicators* are little more than clues for human readers well-versed in one of the two languages and cannot be interpreted automatically. The types of indicators employed are only partially formalizable (examples given here are German to French translations):

- **synonyms**
  eg. **stark** (mächtig) → *puissante*
  gloss: 'strong' in the sense of powerful
- **typical collocations**
  eg. **stark** (Schmerzen) → *violente*
  gloss: 'strong' pains
- **meta-type sem/syn information**
  eg. **stark** (bei Massangabe)
  gloss: 'strong' used with measurements
  example: 2 cm ˜ → *2 cm d'epaisseur*
- **multi-word expressions**
  eg. ˜ Raucher → *grand fumeur*
  gloss: 'strong' smoker (i.e. heavy smoker)

4.2.2 Typical collocations

The information in the entries is rich in both monolingual and bilingual semantic information. The 'semantic indicators' and typical phrases are often indicative of a set collocations (hypernyms, for example, are often used to stand for a class of collocations).

Following is a list of collocations taken from just a few adjective entries.

- **deep** →
  (water, sorrow, thoughts) profond(e)
  (voice) grave
- **hard** → dur(e)
  (work) dur
  (think, try) serieusement
- **high** → haut(e)
  (speed, respect, number) grand(e)
  (price) élevé(e)
  (wind) fort(e), violent(e)
  (voice) aigu(aiguë)

These simple entries are clearly useful for gathering a starting list of collocations, all of them linguistically relevant *l-collocations* (as opposed to an initial set of *p-collocations* found in corpora which would require hand filtering).

The grouping of objects may reflect semantic patterning, though not always, and in fact,

[12] The explicit SGML mark-up also provides the means for displaying the results of a given query as if a printed dictionary had been consulted.

not very often. Support verbs in particular are often listed with coherent semantic classes, eg. under "take" we can find classes like *step, walk* or *effort, courage.* But as the sample translations for **deep** and **high** illustrate, we are often confronted with groups that could hardly be viewed as semantically homogeneous. Nevertheless, the examples given in the entries usually cover the basic sense distinctions and thus provide a useful set to test initial hypotheses. A larger collection of data and more extensive empirical studies may help to identify more general and also more refined classes of regularities that are not apparent in these selected examples.

One immediate way to extend this list of examples is to look for phrase or candidate collocations within the translations, i.e. searching on the word contained in the translation and example fields. Following is a list of examples extracted from a search for adjectives in the translation field (the headword is given in boldface).

- deep voice ← *(voix/son)* **grave**
- deep hate ← *(mépris)* **profond**
- deep/great ← **vif** *(regret/deception)*
- hard choice/problem ← *(choix/problème)* **difficile**
- hard work ← *(travail)* **dur**/ferme
- hard (insensible) person ← *(coeur/-personne)* **sec**
- high price ← *(prix/sommet)* **élévé**
- high returns ← *rendement* **fort**
- high/low tide ← **marée** haute/basse
- high vitamin/mineral/etc content ← **richesse** *en vitamines*
- high blood pressure ← *faire ou avoir de* **la tension**

In comparing this list with the examples cited above (searching on headwords) it is immediately apparent that the information provided is not symmetric. Given the serious space restrictions on these dictionaries in particular, this is not surprising, though a quick look at larger dictionaries will show similar discrepancies.[13] More importantly, what does become apparent in these examples, is the cross-linguistic verification of our definition of collocations. It also substantiates the claim that the nouns (bases) the adjective (collocate) selects have some kind of semantic homogeneity. In essentially all cases (the exceptions seem to represent metaphorical and idiomatic expressions) the nouns retain their regular meaning, the adjectives are selected (and ultimately interpreted) in the contexts of the nouns, which do have a straightforward translation. (cf: observations in section 3.3.1). Given our relatively limited understanding of these phenomena and lack of ability to formalize this system, we can perhaps better appreciate the rather intuitive and incomplete information provided by the lexicographers and hence the assymetries across entries.

4.2.3 Sense distinctions and translation

Another type of information we find in the entries (other than specific collocations) could be interpreted as sense disambiguation information, i.e. when more than one translational equivalent is given without further restrictions, we are usually confronted with at least partial synonyms. In the translation of the German adjective *stark* (strong) from French, for example we find the following translations:

- **farouche** → *(volonté, haine, résistance)* stark, heftig (gloss: heavy or forceful)
- **intense** → stark, intensiv (gloss: intensive)
- **violent** → heftig, stark
- **virulent** → stark, tödlich (gloss: deadly)

The synonyms supplied for each translation choice seem to indicate a possible sense distinction. The underlying basic concept is that of intensification which is precisely the interlingual concept mediated by Mel'čuk's LF MAGN. However, in each case the meaning

[13] Note also that only a subset of the entries containing the words we are interested in have been listed here; fixed expressions and other idiosyncractic examples were deleted.

is refined, as suggested by the glosses supplied above. Following Pustejovsky's notion of *Qualia* [Pustejovsky, 1991], part of the choice for a given collocate will depend on the semantics of the noun, but this must perhaps be judged in interaction with particular senses of a given adjective rather than the word itself. It is unfortunate that translation dictionaries in general do not represent sense distinctions in the translation fields. As these examples suggest, if we are looking to use the lexical functions for generation, we may need to refine them on the basis of the senses in order to choose from the set of possible translations.

A next step would be to look at whether the sense distinctions found in monolingual dictionaries can be of service in helping to classify translation equivalents (cf. the discussion in the following section); another is to look for the semantic structures and primitives provided in thesauri. A search on the Roget Thesaurus [Roget, 1990], again searching from within, i.e. all classes where *strong* is given as example, provided some initial categories. Three major categories which included *strong* as a possible adjective were "quantity by comparison with a standard", "degree of power" and "physical energy". Though the first category does not seem to help in distinguishing the translations given above, the latter two do help in distinguishing *stark* as "intense" and *stark* as "violent". This very preliminary investigation, though far from giving a systematic account of this problem, suggests the potential of combining the information from a number of resources in view of working towards more general accounts of the data.

4.3 Semantic classes and chains

Given the numerous pairs of bilingual dictionaries and the ability to search inside entries, we explored the idea of chaining through entries. Starting with a headword in one language we find a number of translations which in turn serve as the query both as headword or as occurring in a translation field. This work is similar to the "sprouting" mechanism described in [Chodorow and Byrd, 1985] which established chains through entries in order to build up a semantic hierarchy and the subsequent work on "bilingual sprouting" discussed in [Byrd *et al.*, 1987]. In contrast to the work cited above, we did not attempt to find explicit and formalized semantic information (such as +HUMAN), nor to automatically derive taxonomies. The intuition is simply that chaining through words and their translations (from headwords to translations and vice versa) we will discover synonym classes across the languages which can help refine our insights into the *semantic* or *conceptual* differences within and across languages.

In a sample chain beginning with the word **heavy** in English, we find three basic translations in French, i.e. *lourde, gros* and *grand.* The entry for the adjective **lourd** lists only the one translation *heavy* but occurs in the entries for 'close', 'heavy', 'hefty', 'weighty', among others. The entry for **gros** gives as translations *big, large, fat, extensive, thick, heavy* and occurs in 'big', 'broadly', 'fat', 'heavy', 'hefty', 'roughly', etc.

As the example above suggests, after only one step through the chain, we can quickly collect a set of words that clearly share a subset of semantic properties. The differences will have to be accounted for in a systematic way if we are not going to list every possible combination in view of translation and subsequent generation. It is clear that a lexical function such a MAGN will account for any number of collocations containing the words given above, but not why one word is chosen over another.

In the entry for **heavy**, for example, we are also supplied with some contextual information necessary for choosing the appropriate equivalent. I.e. it translates as *gros* in the context of *work, sea, rain, eater* but as *grand* in the case of *drinker* and *smoker.* It is also interesting to note that two of the three main sense distinctions given in OALD correspond approximately to the first two translations, i.e. 1. weight, and 2. size, force, amount. A possible line of investigation would be to see how well these distinctions overlap with the semantic groupings found in thesuri[14].

Another study starting with the adjective **fort** in French and chaining through German and English, yielded the following set of semantically related adjectives.

farouche, intense, puissant, violent, virulent, fort, doué, capable, vigoureux, solide, vif, haut, grand, élevé, aigu, grand, formidable, gros, corpulente, dur, sèrieusement, agressive, bruyant, sonore, beaucoup

Though every native speaker would know precisely which of these adjectives could occur with any given noun, we do not have any even semi-formal account of why this is the case let alone how they will behave in translation. One question to be studied is whether we can identify subsets of this class that behave similarly with respect to collocations. It is this type of data and more that we will need to analyze and test our hypotheses on.

Conclusion

As the title suggests, the main focus of the project "Collocations and the Lexicalisation of Semantic operations" is the semantics involved in collocational constructions. In Mel'čuk's approach, the meaning of each 'collocate' can be identified as an instance of one of the fifty or so Lexical Functions. Whether or not each collocation can be assigned such a function, and whether or not these functions make any sense, is partly a matter of empirical investigation.

In this paper, we have discussed some possibilities in which this empirical investigation can be carried out. Although most dictionaries do not immediately provide the information we are after, they are a rich source of lexicosemantic knowledge which can be exploited in many ways for various purposes. The primary use we have been talking about involves the research on the semantics of collocations rather than the automatic acquisition of lexical information.

For instance, simple chaining through the dictionaries provides us with synonym sets (or some kind of equivalence set) which capture facets of meaning which may point to a lexical function. But one could try to go beyond the simple identification of lexical functions and look for further evidence for generalizations and refinements. Indeed, one of the drawbacks of the lexical function approach as used in the ECD is that each function has to be enumerated completely. In this account, as in many others, collocations are considered to be totally unpredictable combinations. However, looking at collocations one is sometimes struck by the semantic patterns suggested by the data. Again there is a need to find more evidence to be able to tackle such questions as: what kind of regularities are we talking about, to what extent are they there, and how can we account for them?

A second drawback concerns the 'refinement' of the functions. Lexical functions often pick out one semantic facet of their argument to operate on. Mel'čuk himself already makes some provisions, but does not provide a systematic theoretical account of these extensions. One could also turn to other theories of semantic structure such as Pustejovsky's *Qualia* theory. We are then lead to the question how these theories fit the phenomenon we are describing, and again, what kind of evidence we can use to decide on particular analyses.

Collocations are an inspiring source of interesting hypotheses concerning all kinds of lexical semantic issues. Working on them we have realised however, that conjectures need thorough testing. The previous sections have served to identify the different types of lexical resources we can exploit in our study of collocations. Electronic access to various collections of data, i.e. mono- and bilingual dictionaries, and thesauri provide the basis for a more in-depth exploration of the phenomena. A next step is to refine the methods for picking out relevant data and combining the information gained from different perspectives.

[14] A simple program to extract phrases from Roget identified ca. 20 expressions containing heavy as an adjective.

References

[Bloksma *et al.*, 1992] Laura Bloksma, Dirk Heylen, and R. Lee Humphreys. Characterisation. Deliverable ET-10/75 [C.1], Collocations and the Lexicalisation of Semantic Operations, Utrecht, July 1992.

[Byrd *et al.*, 1987] R. Byrd, N. Calzolari, M. Chodorow, J. Klavans, and M. Neff. Tools and methods for computational linguistics. *Computational Linguistics*, 13(3-4):219–240, 1987.

[Calzolari, 1983] N. Calzolari. Semantic links and the dictionary. In *Proceedings of the Sixth International Conference on Computers and the Humanities*, pages 47–50, Maryland, 1983.

[Chodorow and Byrd, 1985] M. Chodorow and R. Byrd. Extracting semantic hierarchies from a large on-line dictionary. In *Association for Computational Linguistics proceedings*, pages 299–304, Chicago, 1985.

[Church and Hanks, 1989] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information and lexicography. In *27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, 1989.

[Collins, 1990] Collins, editor. *Collins French-English English-French Dictionary, Collins German-English English-German Dictionary, Collins French-German German-French Dictionary.* Collins, London, Glasgow, 1990.

[Heylen *et al.*, 1992] Dirk Heylen, Kerry G. Maxwell, Tim Nicolas, and Susan Warwick-Armstrong. Analysis of existing dictionaries. Deliverable, Collocations and the Lexicalisation of Semantic Operations, Utrecht, July 1992.

[Heylen, 1992] Dirk Heylen. Lexical functions and knowledge representation. In *Proceedings of the Second International Workshop on Computational Lexical Semantics*, Toulouse, 1992.

[Hornby, 1974] A.S. Hornby. *The Oxford Advanced Learner's Dictionary of English.* Oxford University Press, third edition edition, 1974.

[Mel'čuk *et al.*, 1984] Igor Mel'čuk, Nadia Arbatchewsky-Jumarie, Léo Elnitsky, Lidija Iordanskaja, and Adèle Lessard. *Dictionnaire explicatif et combinatoire du français contemporain.* Les Presses de l'Université de Montréal, Montreal, 1984.

[Petitpierre and Robert, 1991] D. Petitpierre and G. Robert. Dico - a network based dictionary consultation tool. In *SGAICO proceedings*, Neuchatel, 1991.

[Pustejovsky, 1991] James Pustejovsky. The generative lexicon. *Computational Linguistics*, 17(4), October-December 1991.

[Roget, 1990] Roget. Roget's thesaurus, 1911, 1990.

[Sanfilippo *et al.*, 1992] A. Sanfilippo, T. Briscoe, A. Copestake, Maria Antonia Martì, Mariona Taulé, and Antonietta Along. Translation equivalence and lexicalization in the acquilex lk. In *Proceedings of TMI 4*, pages 1–11, Montreal, 1992.

[Smadja, to appear] Frank Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, to appear.

[Vossen, 1991] P. Vossen. Comparing noun-taxonomies cross-linguistically. ESPRIT BRA-3030 ACQUILEX Working Paper Number.014, English Department, University of Amsterdam, 1991.

[Wall and Schwartz, 1990] L. Wall and R. L. Schwartz. *Programming Perl.* O'Reilly & Associates Inc., Sebastopol, CA., 1990.