Dirk Heylen
OTS, Universiteit Utrecht

Kerry Maxwell
CL/MT Group, Essex University

Lexical Functions and the Translation of Collocations

Abstract

This paper discusses the lexicographical concept of *Lexical functions* (Mel'chuk and Zholkovsky, 1984) and their potential exploitation in the development of a machine translation lexicon designed to handle collocations. We show how lexical functions can be thought to reflect crosslinguistic meaning concepts for collocational structures and their translational equivalents, and therefore suggest themselves as some kind of language–independent semantic primitives from which translation strategies can be developed.

1. Introduction

Collocations present specific problems in translation, both in human and automatic contexts. If we take the construction *heavy smoker* in English and attempt to translate it into French and German, we find that a literal translation of *heavy* yields the wrong result, since the concept expressed by the adjective (something like 'to excess') is translated by *grand* in French and *stark* in German. The example is a good illustration of the fact that languages differ in how they express the concepts mediated in collocational structures, e.g. a *heavy smoker* in English is a 'large' smoker in French *(grand fumeur)* and a 'strong' smoker in German *(starker Raucher)*. We observe then that in some sense the adjectives *stark, grand* and *heavy* are *equivalent* in the collocational context, but that this is of course not typically the case in other contexts, cf: *grande boîte, starke Schachtel* and *heavy box,* where the adjectives could hardly be viewed as equivalent. It seems then that adjectives which are not literal translations of one another may share meaning properties specifically in the collocational context.

How then can we specify this special equivalence in the machine translation dictionary? The answer seems to lie in addressing the concept which underlies the union of adjective and noun in these three cases, i.e. intensification, and hence establish a single meaning representation for the adjectives which can be viewed as an interlingual pivot for translation. The CEC project ET–10/75: Collocations and the Lexicalisation of Semantic Operations, is concerned with investigating the *lexical functions* (LF's) of Mel'chuk (Mel'chuk and Zholkovsky 1984), as a candidate interlingual device for the translation of adjectival and verbal collocates. In this paper we

will attempt to provide an overview of the key areas of research entailed by the project, including a characterisation of collocational structures, an evaluation of Lexical Functions and proposals for representation and translation strategies.

2. A characterisation of collocational structures

In Mel'chuk's 'Explanatory Combinatory Dictionary' (ECD, see (Mel'chuk et al., 1984)) expressions such as *une ferme intention, une résistance acharnée, un argument de poids, un bruit infernal* and *donner une leçon, faire un pas, commetre un crime* are described in the lexical combinatorics zone. These 'expressions plus ou moins figées" are considered to consist of two parts – which we will call the *base* and the *collocate.* In the examples above the nouns are the bases and the adjectives and the verbs are the collocates. The idea that all adjective collocates and all the verb collocates share an important meaning component – roughly paraphrasable as *intense* and *do* respectively – and the fact that the adjectives and verbs are not interchangeable but are restricted with this meaning to the accompanying nouns, is coded in the dictionary using lexical functions (in this case **Magn** and **Oper**).

This class of loosely fixed combinations can be identified as collocations. We try to determine some properties that fit this set of expressions and that are compatible with a number of ways the term collocation is used in the literature.

Looking at several characterisations,[1] we see the following predicates surface repeatedly, albeit in different guises: *recurrent, idiomatic, contextually restricted, cohesive* and *arbitrary.* These notions can be summed up as 'collocations are cohesive, recurrent, arbitrary combinations of words which are not idioms but in which the (figurative) meaning of one part is contextually restricted to the specific combination".

Let us briefly explain these notions and see which properties we have used in our own characterisation. The first notion, *recurrence* is most typically captured in the definition of collocation as a 'recurrent combination of words that co–occur more often than expected by chance" (Smadja 1993). We have not included this notion in our definition, although we have used it to extract potential collocations from texts (see (Heylen et al., 1993)). It is a surface characteristic related to the occurrence of collocations.

The second notion pertains to the semantic properties of the parts making up the collocation and the way they are combined. In the ODCIE (Cowie and Mackin 1975), it is stated that a collocation is 'not an idiom because the meaning of the whole reflects the meaning of the parts".[2] Mel'chuk also says that they are not 'idioms stricto sensu". As far as the meaning of the parts is concerned it is said that 'one word has a figurative sense", but 'the other element appears in a familiar, literal sense". This second notion is not found in every definition of collocations. In most corpusbased approaches to

collocations, for instance, this property is absent. It is not mentioned in the (Benson et al. 1986) definitions either. Another problem is that it is sometimes difficult to decide whether we are dealing with a figurative meaning or with a meaning which is not the primary meaning. The least we can say is that in most cases the meaning of the collocate is not its most prominent one.

Another way to say that this meaning is special, is by using the third notion. The particular reading of the collocate is one that is 'not found outside that limited context" (ODCIE). This means that one part is only used in this sense in restricted contexts (although some lexical variation is possible). Also the term 'loosely fixed combination" covers this property. It also ties in with the fourth notion, *cohesiveness* which means that 'the presence of one or several words of the collocation often implies or suggests the rest of the collocations". (Smadja, also BBI). We combine these notions by stating that the base of a collocation selects a specific word (or a limited set of words) to express a certain meaning and this selected word, the collocate, is only used with a limited set of bases to express this meaning. This combination of properties is responsible for the cohesion between the elements and probably also for their recurrence (certainly if we assume that the meanings are ones that are often expressed).

The fifth notion 'arbitrary' (Benson, Smadja) is sometimes also referred to as 'lexical'. There is no semantic reason why a smoker is 'heavy' in English and 'strong' in German. This is to a large degree determined by coincidental lexical selection.

Our notion of collocation is a combination of these properties (though it considers the first as a derived property rather than a criterial one), and adds the idea that to a large extent the figurative meaning of the part can be identified with lexical functions.

3. Analysis of lexical functions

In this section we provide a brief introduction to Mel'chuk's proposals concerning lexical functions (Mel'chuk et al., 1984), (Mel'chuk and Polguère 1987), (Mel'chuk and Zholkovsky 1988a). Broadly speaking, Lexical Functions (henceforth LF's) are used to describe systematically certain semantic and collocational relations existing between lexemes. They apply at the deep syntactic level of the Meaning–Text Model (MTM),[3] and are used to indicate either a set of phraseological combinations related to a keyword (argument lexeme to which they apply) or those words which can replace a keyword under certain conditions. We focus on the former class, called syntagmatic LF's.

A definition of the notion of lexical function can be found in (Mel'chuk and Zholkovsky 1988b: 51): 'A LF is a function in the mathematical sense representing a certain extremely general idea, such as 'very', 'begin', or 'implement', or else a certain semantico–syntactical role. A lexical function

f associates with a word W_0 called its argument, or KEY WORD, the set of words and phrases which express – contingent on W_0 – the meaning or role which corresponds to **f**.'

We should note that the co–occurrence restrictions these lexical functions are intended to capture are those that are truely linguistic and not the restrictions between lexemes that 'cannot co–occur only because of their meanings and of our knowledge of the world'.[4] So the philosophy is that LF's are in no way intended to refer to the semantics of the lexemes over which they operate. We should note however that the relation denoted by *the LF itself* is somehow semantic in nature, cf: the LF **Magn** denoting 'intensification' and the LF **Degrad** denoting a process of 'becoming worse or bad'.

4. Issues in translation

Despite their differences, the classical MT *transfer* and *interlingua* architectures share the same basic mechanics. Both fall within the paradigm that is concerned with mappings between symbolic representations that start and end with natural language. Differences between these architectures and their variants can be characterised by the number of representation levels, their interpretation, and the mappings between them. In a transfer architecture an expression in the source language will be analysed up to some specific level of representation. Next, the resulting structure is mapped onto a similar structure of the target language by specific transfer rules. From this structure the target language expression is generated. Interlingua systems assume a level of representation with structures that are shared by both the source and the target language. In this case no transfer mapping is necessary, or one could say that the transfer mapping is the identity function.

The project has tried to investigate the use of Lexical Functions as an interlingual device, i.e. one which is shared by the semantic representations of collocations in the language pairs.[5]

The typing of a collocation with such a function opens up the way to a treatment of collocations inside a given language module and hence to a substantial reduction in the number of collocations explicitly handled in the multilingual transfer dictionary. The existence of a collocation function is established during analysis. This information is used to generate the correct translation in the target language. To illustrate, the English analysis module might analyse (1) *heavy smoker* as (2) **Magn**(smoker). The transfer module maps (2) onto (3) **Magn**(fumeur) which is then synthesised by the French module to (4) *grand fumeur.*

The example points out that the translation strategy is a mixture of transfer and interlingua. The bases are transferred but the representation of the collocate is shared between the source and the target representation. This treatment of collocations rests, among others, on the assumptions that there are only a limited number of lexical functions, that lexical functions can be

assigned consistently, that all (or a significant number of) collocations realise a lexical function, that lexical functions are not restricted to particular languages, etc. In the following paragraphs we discuss two problems. The first deals with the appropriateness of Lexical Functions as an interlingual device. The second is concerned with the problems that arise when collocations do not translate into collocations.

4.1 Lexical functions as interlingua

4.1.1 Overgenerality

An important problem stems from the interpretation of LF's implied by their use as an interlingua – namely that *the meaning of the collocate in some ways reduces to the meaning implied by the lexical function.* This interpretation is trouble–free if we assume that LF's always deliver unique values; unfortunately cases to the contrary can be readily observed. An example attested from our corpus was the range of verb adverb constructions possible with the verbal head *oppose,* e.g. *adamantly, bitterly, consistently, steadfastly, strongly, vehemently, vigorously, deeply, resolutely.*

The function **Magn** is an appropriate descriptor in all cases since each adverb functions as a typical intensifier in this context. However each adverb also denotes some other meaning aspect(s), e.g. *consistently* suggests something like 'continuingly', *bitterly* suggests 'animosity', etc. These meaning aspects are not captured by the function, i.e. **Magn** refers to the intensification device inherent in each adverb but cannot say how these possible intensifiers differ. Now, one may argue that the difference between these adverbs can be highlighted in the semantic specifications (or definitions) associated with their individual entries in the lexicon. This does not, however, offer any consolation if our aim is to exploit purely LF's as an *interlingual* device at the point of translation. Their imprecision will mean that we have no means of distinguishing between the various intensifiers possible in the context of a given keyword, and hence will not have sufficient information to choose the correct translation where, correspondingly, multiple possibilities exist in the target language.

4.1.2 Possible enhancements

It is essentially in addressing the issue of overgenerality that Mel'chuk introduces sub– and superscripts to lexical functions, enhancing their precision and making them sensitive to meaning aspects of the lexical items over which they operate. Superscripts are intended to make the meaning of the LF more precise, subscripts are used to reference a particular semantic component of a keyword. The introduction of such devices into the account of LF's demonstrates both the need for precision and the fact that it does seem necessary to address semantic aspects of lexemes standing in co–occurrence relations. In fact it has been suggested by some (e.g. (Anick

and Pustejovsky 1990), (Heid and Raab 1989)) that collocational systems may be systematically predictable from the lexical semantics of nouns.

In an attempt to explore this notion further, we have investigated the approach to nominal semantics known as *Qualia* structure (Pustejovsky 1991) and considered how this may complement the LF notion to improve its descriptive power.[6] Among the promising avenues that occur to us are, firstly, the postulation of LF subscripts based on the four Qualia roles (assuming that these are the lexically most relevant aspects of noun semantics) and, secondly, the application of LF's to semantic (Qualia) structures rather than monolithic lexemes; e.g. the LF **Bon** is used in delivering evaluative qualifiers which are standard expressions of praise/approval. One could imagine application of the function over the Constitutive and Agentive roles of the noun *lecture,* to deliver: – **Bon**(Const: lecture) = informative and **Bon**(Agent: lecture) = clear.

The Constitutive role refers to constituent parts of something; in this case, one could conceive its interpretation as the *contents* of the lecture. The Agentive role refers to the 'bringing into existence' of the noun. A *speaker* is clearly involved here so the value refers to a positive attribute of speech.

In both cases the idea is that the precision of the lexical function is essentially enhanced by appealing to the semantic facets of its argument.

4.2 Syntactic divergences

A thorough analysis of the translational patterns exhibited by collocations reveals, inevitably, that there are cases which do not fit our basic assumptions about structural identity and lexical mismatches across language pairs. In other words, there are examples of collocations whose translations do not count as collocations according to our criteria. If we are to maintain a consistent interlingual approach to the translation of these cases, we must revise (extend) our LF–based approach accordingly, as we now discuss.

4.2.1 Compounding and lexicalisation

Crosslinguistic analysis reveals many cases where nominal–based collocational constructs are realised as compounds in Germanic languages, e.g. bunch of keys –> sleutelbos.

The orthodox perception of LF's is that they embody *inter–* rather than *intra–*word relations, which implies that they cannot offer a description of the processes underlying word (or compound) formation. However a possible account of such phenomena may be developed from the concept of *merged* LF's cf: (Mel'chuk and Zholkovsky 1970). Merged LF's are intended to be used in cases where a value lexeme exists which appears to effectively reduce ("merge') an LF meaning and its specified argument to a single lexicalised form, rather than projecting a syntagmatic unit. We could argue that in cases of compound formation, exactly the same process is to be accounted for,

since the compound embodies both the concept mediated by the LF and its argument lexeme. We could therefore allow compounds to be delivered as values of merged LF's, e.g. //**Mult**(sleutel)= sleutelbos.

We can effect a mapping between merged and unmerged LF's and therefore capture the correspondence between distinct structural realisations of the same concept through the use of Mel'chuk's *lexical paraphrasing rules*. For instance, one could conceive of a lexical paraphrasing rule as follows: – W + **Magn**(W) ↔ //**Magn**(W).

Further examples exist where productive morphological processes (e.g. affixation) lead to the lexicalisation in one language of concepts that exist as syntagmatic constructs in another. Again, we suggest the use of merged LF's and corresponding mappings via lexical paraphrasing rules as a possible translation strategy in these cases.

4.2.2 'Literal' translation

In some cases collocations are translated as ordinary 'literal' expressions. The regular existence of such translations would cause problems for our account if we were to assume that LF's were only to be used for Adjective–Noun combinations which count as collocations on some 'linguistic' basis connected with the figurative nature of the collocate. LF's could therefore not be used to describe such combinations as these and our interlingual mapping would break down. The answer may be to allow LF's into the domain of such combinations. The problem here is that we throw wide open the issue of overgenerality of LF's; if we allow LF's to apply to AN combinations of all kinds then the potential range of values associated with the functions will increase explosively. This issue remains a problem, though it may be alleviated by the postulation of 'default' values for certain LF's which could be overridden where lexically specific values (collocates) exist.

5. Summary and conclusions

In this paper we have discussed how the lexicographical concept of *lexical functions* introduced by Mel'chuk to describe collocations, can be used as an interlingual device in the machine translation of such structures.

Our use of lexical functions as an interlingua assumes that the relevant aspects of the meaning of the collocate are fully captured by the LF. The LF therefore determines the accuracy of translations, which may be impoverished due to the generalised nature of basic LF's. We have suggested some ways in which LF's can be enriched with lexical semantic information to improve translation quality.

The interlingua level abstracts away from specific syntactic realisations. Given that collocations may translate as non–collocations, we have to also

provide a way to represent these expressions using lexical functions. We have provided some suggestions on how to proceed in such cases.

Notes

1 In particular, (Cowie and Mackin 1975), (Mel'chuk et al., 1984), (Benson et al. 1986), (Cruse, 1986), (Church and Hanks 1989), (Smadja 1993). See also (Heid et al. 1991).
2 In (Cruse 1986) it is stated that these expressions are "fully transparant in the sense that each lexical constituent is also a semantic constituent".
3 LF's appear in the Explanatory Combinatorial Dictionary (ECD), the lexical component of the MTM.
4 Note that those co–ocurrence properties which reference the denotations of lexemes are the domain of so–called **selectional restrictions**. Here a specific aspect of the semantics (denotation) of the nouns is referenced.
5 For another application of LF's in a multilingual NLP context see (Heid and Raab 1989). For other treatments of collocations in language generation see (Nirenburg et al. 1988) and (Smadja and McKeown 1991).
6 For a comparison between aspects of Qualia structures and lexical functions see (Heylen, to appear).

References

P. Anick and J. Pustejovsky 1990. An application of lexical semantics to knowledge acquisition from corpora. In *Coling*, Helsinki.
M. Benson, E. Benson, and R. Ilson 1986. *The BBI Combinatory Dictionary of English.* John Benjamins, Amsterdam.
K. Church and P. Hanks 1989. Word association norms, mutual information and lexicography. In *ACL*, Vancouver.
A.P. Cowie and R. Mackin 1975. *Oxford dictionary of Current Idiomatic English.* OUP, London.
D.A. Cruse 1986. *Lexical Semantics.* CUP, Cambridge.
U. Heid and S. Raab 1989. Collocations in multilingual generation. In *EACL*,: 130–136, Manchester.
U. Heid, W. Martin, and I. Posch 1991. Feasibility of standards for collocational description of lexical items. *ET7.*
D. Heylen, K. Maxwell, and S. Armstrong Warwick 1993. Collocations, dictionaries, and mt. In *Building Lexicons for Machine Translation*, Stanford. AAAI Spring Symposium.
D. Heylen to appear. Lexical functions and knowledge representation. In P. Saint–Dizier and E. Viegas, editors, *Computational Lexical Semantics.* CUP.
I.A. Mel'chuk and A. Polguère 1987. A formal lexicon in the meaning–text theory. *Computational Linguistics*, 13(3–4).
I.A. Mel'chuk and A.K. Zolkovsky 1970. Sur la synthèse sémantique. T.A. *Informations*, 2:1–85.
I.A. Mel'chuk and A.K. Zolkovsky 1984. *Explanatory Combinatorial Dictionary of Modern Russian.* Vienna.
I.A. Mel'chuk and A.K. Zolkovsky. The explanatory combinatorial dictionary. In M. Evens, editor, *Relational Models in the Lexicon*,: 41–74. CUP, Cambridge.
I.A. Mel'chuk and A.K. Zolkovsky. The explanatory combinatorial dictionary. In M. Evens, editor, *Relational Models of the Lexicon.* Cambridge University Press.
I.A. Mel'chuk, N. Arbatchewsky–Jumarie, L. Elnitsky, L. Iordanskaja, and A. Lessard 1984. *Dictionnaire explicatif et combinatoire du français contemporain.* Les Presses de l'Université de Montréal, Montreal.
S. Nirenburg, R. McCardell, E. Nyberg, S. Huffman, E. Kenschaft, and I. Nirenburg 1988. Lexical realization in natural language generation. In *Theoretical and Methodological Issues in MT of Natural Languages*, Pittsburgh.
J. Pustejovsky 1991. The generative lexicon. *Computational Linguistics*, 17(4).
F. Smadja and K. McKeown 1991. Using collocations for language generation. *Comput. Intelligence*, 17.
F. Smadja 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.