Nicolas DUFOUR, DEFI, CIPL – University of Liège

Recognizing collocational constraints for translation selection: DEFI's combined approach

Abstract

This paper describes an experiment carried out in the framework of the DEFI word sense discrimination project. That experiment aims at determining the strengths and weaknesses of the three methods we use in order to establish a semantic link between collocates found in unrestricted text and collocational constraints imposed on translations in our bilingual dictionaries. Establishing such links is shown to be vital for context-sensitive translation selection.

1. Introduction

DEFI's objective is to create the prototype of an 'intelligent' dictionary look-up program that would provide the reader of a text in a foreign language (in this case English) with the most appropriate translation (into French) of any word he/she selects online (see Michiels 1996). The intended prototype, an early version of which is already up and running, can thus be regarded as a 'comprehension assistant' similar in its goals to Rank Xerox's LOCOLEX (Bauer *et al.* 1995), albeit with a different, more semantically-driven approach.

Our look-up program (or 'text-dictionary matcher')[1] processes the user-selected word in two ways:

- In a first stage, the matcher checks whether the word might be part of a larger multi-word unit (MWU). This implies comparing the structure of the source sentence around the selected word with that of the MWUs containing that word in our bilingual dictionary, a process that relies on subtle and constantly modifiable heuristics. Once a valid MWU has been identified, however, translating it is mostly a straightforward affair since most MWUs (phrasal and prepositional verbs excepted) have only one or very few senses.
- If the selected word does not appear to belong to an MWU, however, the matcher often finds itself facing a highly polysemous single-word lexeme with a range of translations climbing to double digits. This is where some kind of semantic analysis of the context has to be performed if we want to avoid providing the user with a lengthy list of translations he/she could have found just as easily by querying any paper or on-line dictionary 'by hand'.

The English-French dictionary we use for translation selection, DEFIDIC, combines the machine-readable versions of two well-known general-use bilingual dictionaries, the Oxford-Hachette (OH) and Collins-Robert (CR) English-French dictionaries. We obtained from the publishers the files used for printing the paper versions of the two dictionaries, a typesetting tape in the case of CR and an SGML-tagged file for OH. The two files were turned into a common format, then merged into a single machine-tractable dictionary (MTD).

Like most modern bilingual dictionaries, both OH and CR make extensive use of collocational restrictions in order to help their user select the correct translation of a word. Colloca-

tional restrictions are actually a more powerful tool than sense restrictions, since the translation of a word depends not only on its meaning but also on the organisation of the lexicon in the target language. Consider the following collocational restrictions and the corresponding translations in CR, all applying to *bright* in the sense of 'shining, full of light' (as opposed to 'cheerful' or 'intelligent'):

> **bright** (*shining*): *eyes* brillant, vif; *star, gem* brillant; *light* vif; *fire* vif, clair; *weather* clair, radieux; *sunshine* éclatant; *day, room* clair; *colour* vif, éclatant, lumineux; *metal* poli, luisant.

Provided with such a wealth of collocational information, a human user can easily choose a translation according to the noun he/she finds associated with *bright* in the text. DEFI's matcher does it as well, comparing collocates attached to the selected word in the source text with those listed in DEFIDIC. Thus, in a sentence like *I love cycling in the bright sunshine of summer days*, *bright* will easily be translated as *éclatant*. The snag, however, lies in the fact that collocational restrictions in the dictionaries are not exhaustive, and that the collocates listed there must be interpreted as thesauric heads rather than as specific lexemes (see also Fontenelle (1997a)). *Gem*, clearly, stands for all kinds of gems and precious stones, just like *metal* stands for all potentially shiny metals. While a human user understands this and makes the necessary adjustments instinctively, automatic look-up systems are bound to fail in the absence of the necessary semantic information. Left to its own devices and relying solely on DEFIDIC, the DEFI matcher would not even be able to translate *bright* in *I love cycling in the bright sunlight of summer days*.

The present paper describes and compares three methods used by DEFI to provide our matcher with the semantic information needed to find out, say, that *sunlight* is just like *sunshine*, or that *copper* and *steel* are just two kinds of *metal*. The three methods are already implemented with some success by our matcher, but a comparative and extensive analysis was lacking up to now. Note that, for pragmatic reasons only, the experiment described here applies only to nominal collocates. Nouns account for over 94% of all collocational restrictions in DEFIDIC, and they are also the part of the lexicon whose semantic organization is the most complex.

2. Three methods for matching collocates

2.1. WordNet query

WordNet (Miller 1990) is a lexical-semantic database organizing the lexicon into sets of synonyms or very near synonyms (hence the name *synset* for the basic group of word forms in WordNet). These *synsets* are linked to one another by the classical semantic relations of antonymy, hypernymy, hyponymy, meronymy and holonymy. The advantage of such a formalized man-made database is obvious: it should (and does) tell us that *sunlight* is a synonym of *sunshine*, hence solving our cycling problem, or that *evening* is a part of *day*. Note that even a relation of antonymy between two words is valuable: antonyms often share semantic properties and collocational constraints, so that a bright *night* is 'bright' in the same sense (and with the same translation) as a bright *day*.

When trying to establish a link between the collocate found in the source text (txtcoll) and a collocate listed in DEFIDIC (diccoll), our look-up program searches WordNet via the synset

references of the two words. The Prolog routines we use for that search are slightly modified versions of the predicates available on the Web site of Princeton University. At the end of the search, the link between the two words is granted a score depending on the kind of relations holding between them. The value of the different scores has been chosen empirically, and is bound to be revised as work and evaluation progress. Here is an overview of the WordNet scores as of now:

Diccoll and txtcoll are synonyms	100
Diccoll and txtcoll are antonyms	40
Diccoll is a hypernym of txtcoll 1 level up	30
Diccoll is a hypernym of txtcoll 2 levels up	15
Diccoll is a hypernym of txtcoll 3 levels up	9
Txtcoll is a hypernym of diccoll 1 level up	10
Txtcoll is a hypernym of diccoll 2 levels up	7
Txtcoll is a hypernym of diccoll 3 levels up	3
Txtcoll and diccoll have a common hypernym	6

Note that the various scores granted to hypernymy are not symmetrical: while *copper* inherits all the properties of *metal* (hence the high score), the validity of the link would be less certain if we had *copper* in the dictionary and *metal* in the source text.

2.2. Category sharing in Roget's thesaurus

Since its first publication in 1852, Roget's *Thesaurus of English Words and Phrases* has become one of the most famous lexicographic works in the English-speaking world. Like any thesaurus, it stores words not according to alphabetical order but according to conceptual similarities — i.e., it groups words exactly the way the DEFI matcher needs to get them in order to match textual and dictionary collocates. Roget's thesaurus divides concepts into six *classes*, each class is divided into *sections*, and each section is divided into *heads* (which we call 'categories'). While classes and sections are much too large and vague for our purposes, the 1000 *categories* of our version of Roget's[2] are lists of terms sharing a common, relatively precise concept — and can thus, with some degree of certainty, be regarded as semantically related. Consider the following 'purity' category as it appeared in our original Roget's file:

```
#960. Purity. -- N. purity; decency, decorum, delicacy; continence, chastity,
honesty, virtue, modesty, shame; pudicity[obs3], pucelage[obs3], virginity.
     vestal, virgin, Joseph, Hippolytus; Lucretia, Diana; prude.
     Adj. pure, undefiled, modest, delicate, decent, decorous; virginibus
puerisque[Lat];
simon-pure; chaste, continent, virtuous, honest, Platonic.
     virgin, unsullied; cherry [coll.].
     Phr. "as chaste as unsunn'd snow" [Cymbeline]; "a soul as white as
heaven"
[Beaumonth & Fl.]; "'tis Chastity, my brother, Chastity" [Milton]; "to the
pure
all things are pure" [Shelley].
```

Of course a lot of (automatic) editing had to be done before Roget's could be used as a machine-tractable thesaurus. Phrases and style tags (`[obs]` for 'obsolete', etc) were discarded, and the whole thesaurus was re-written into a list of Prolog structures indicating, for each word, where it appears in the original file. Each category is divided into part-of-speech sections, which are themselves divided into first- and second-level sublists (according to their separators: full stop or semi-colon).

When trying to match txtcoll and diccoll in Roget's, the DEFI matcher compares the exact references of all occurrences of the two words within the thesaurus. No link is established unless the two candidates share at least a category number and a part of speech. The score assigned to their relation is 10 if they share only these, 20 if they belong to the same first-level sublist and 30 if they appear in the same second-level sublist. In the category reproduced above, for instance, the link between *decency* and *delicacy* would be worth 30, that between *decency* and *continence* would be worth 20 and that between *decency* and *vestal* would be worth 10. Note that Roget's scores are cumulative: if two words meet at different places in the thesaurus, the scores assigned to their various co-occurrences are added. This explains, for instance, that the *flame/fire* pair receives a score of 70 or the *pressure/weight* pair a score of 60.

2.3. Metalinguistic slot sharing in DEFIDIC

Metalinguistic slot sharing (MSS), an idea that was first described in Montemagni *et al.* (1996), is by far the most 'intuitive' method used by DEFI for comparing textual and dictionary collocates. In their 1996 paper Montemagni *et al.* argue that it is possible to establish the conceptual relatedness of two words using the metalinguistic information provided by the dictionaries themselves, and more specifically by comparing the contents of their collocate lists. The basic assumption underlying the MSS approach can be summed up as follows: *two words that appear alongside each other in the same collocate list (or 'metalinguistic slot') are likely to share certain semantic properties.* The nature of that 'similarity' is impossible to determine automatically, and often falls outside the relationships usually taken into account by such lexical databases as WordNet. Consider for example *room* and *day* in the collocate lists of *bright* above. *Room* and *day* are related neither by synonymy nor antonymy, nor hyper/hyponymy, nor a part-whole relation, yet they are similar in that they can both be said to be *bright* — and with the same sense and translation of *bright*.

Of course the co-occurrence of two words in a single collocate list is insufficient evidence that these two words are closely related. An efficient implementation of MSS requires the highest possible number of collocate lists, which allows to take into account the frequency of co-occurrence of two words. The database we use for MSS computation was derived automatically from DEFIDIC, which, as the combination of two collocate-rich dictionaries, goes some way towards quenching MSS's thirst for data. DEFIDIC boasts a total of 139,996 collocates spread over 79,967 lists, from which were extracted 37,959 multi-collocate lists numbering a total of 100,988 collocates. Slots featuring only one collocate, of course, are of no use for MSS computation.

DEFI's MSS database is made up of a list of Prolog structures providing, for each DEFIDIC collocate, a list of codes referring to all the multi-collocate metalinguistic slots in which it appears. Once it has been provided with that data, the DEFI matcher simply has to compute the intersection of two lists in order to determine the MSS score of two words. A score of 4 is assigned for each shared slot, a relatively small value similar to the worst possible score of a 'successful' WordNet query.

2.4. Combining strategies

The DEFI matchers, when confronted with a textual collocate it cannot find among those listed in the dictionary, tries to match them through the three methods consecutively. The score that is finally assigned to each txtcoll/diccoll pair is the sum of the three results, and that score is added to other marks gathered by a given translation for other reasons. Although computationally very heavy, this method is clearly the best way to exploit the complementary strengths of the three approaches. Note that whereas even the lowest score is used by the DEFI matcher regardless of its likely relevance — assuming that, if some other translation is more appropriate, its collocates will match better and win the day anyway —, for the separate analysis proposed here a global score of 20 must be regarded as a threshold for any match to be considered conclusive.

3. Choice of test sentences

The analysis described here was centered on the dictionary collocates of five highly polysemous words chosen precisely for the high number of collocates accompanying their translations in DEFIDIC: *bright*, *clear* (adj and vt), *cut* (adj and vt), *heavy* and *light* (adj and vt). For each of these words around 90 collocates were extracted from the British National Corpus (BNC) and matched with the collocates listed in the dictionary. All successful matches (i.e., showing clear results, even if wrong) were then listed separately, showing clearly the scores achieved by each of the three matching methods.

While collocational restrictions in DEFIDIC have been chosen by the OH and CR lexicographers for their perceived 'typicality', one should bear in mind that, in the not-so-brave world of English usage, words do not always collocate in a 'typical' fashion — which does not prevent a human user from using bilingual dictionaries more or less successfully. The test sentences used in this experiment have therefore been extracted half-randomly, making sure simply that they did include a collocate attached to one of the five test words (either as noun modified by an adjective or as object of a verb) and that this collocate was not to be found 'as such' in the relevant DEFIDIC entries. As an illustration of the randomness of the samples, consider the following test collocates, each time the first of their series:

bright: firelight, tooth, glow, orange, idea, beam, window, birdsong, schoolboy
clear (adj): advantage, picture, fluid, tone, print, benefit, effect, statement, philosophy
clear (vt): tile, complexion, name, smell, passage, dropping, muck, waste, fluke, blockage
cut (adj): diamond, piece, fringe, panel
cut (vt): leg, tip, sickle, tree, cake, service, bill, trouser, amount, chicken, shoelace, space, loss
heavy: cloth, usage, good, curtain, flow, garment, cost, spender, price, vehicle, disgust, breast
light (adj): railway, beam, blue, moisturising, touch, relief, stroke, supper, dish, tap, sleep

Note that the average collocates were actually less 'typical' than these, since the most typical collocations are bound to be more frequent — and thus to appear first.

4. Results

Assessing the results of such an analysis is difficult indeed, because the perceived 'quality' of a match:
a) depends on the subjective appreciation of the tester;
b) can never be expressed by a simple mark.

Nevertheless, and with all due caution, all results were assigned a mark ranging from 1 to 5. 1 and 2 indicate very poor or just tendentially poor results, 3 indicates an undecided match (no score as high as the threshold of 20), and 4 and 5 indicate good or very good/perfect matches. The following, somewhat sobering percentage ratios were obtained:

1	2	3	4	5
3.5	18.2	42	7.1	29.2

Although a success rate of just over 36% definitely looks bleak, one should never forget that the experience was carried out using worse-than-randomly chosen collocates of highly polysemous words — the author was undoubtedly sticking his neck out. After all no present-day full-scale NLP system would be able to determine the sense of *bright* in such contexts as *a bright patch on her belly* or *maybe herons are bright enough to realise that [...]* or *bright threads running through our Christian living*. In many cases, the immediate context (just one collocate) simply isn't enough, and context-sensitive word sense discrimination requires a degree of world knowledge and a message understanding capacity which no NLP system possesses on such a scale. Only a human user knows that a bright *sky* is a sky without clouds, and thus not to be related to a bright *cloud* — although *sky* and *cloud* are, at first glance, close parents.

Furthermore, the uncertainty and failure rates would actually be considerably lower in the output of the main DEFI matchers: many 'uncertainties' were actually correct results, and mistakes in collocate matching do not necessarily result in incorrect translations.

It appears clearly that collocate-matching, with a few lucky exceptions, is possible only when the two items being considered really belong to a set of semantically related words — in other cases, even case-hardened human translators would be hard put to choose for one single translation. Attachments are much better in more 'favourable' situations, as illustrated in the following successful attachments of *bright* collocates:

> firelight, glow, beam, sunlight, spark, illumination, gloss, flame, flash ↔ *light*
> orange, coloration, pink, shade ↔ *colour*
> idea, writer, thing (figuratively, in *bright young things*), man, woman, people ↔ person
> schoolboy, lad, boy, pupil ↔ child
> evening, night, morning ↔ day

By examining more closely the attachments performed by the collocate matcher, it is possible to make out the strengths and weaknesses of each method — and to note, from the outset, that they are complementary.

WordNet has the sturdy reliability of a man-made, state-of-the-art lexico-semantic database. WordNet is very good at providing synonyms, antonyms and hypo/hypernyms, and very high scores are rarely achieved without some contribution from its part. More intuitive conceptual matches, however, cannot be expected, and exotic terms are not its domain either. Just like the two other databases, WordNet suffers from the ambiguity of its own constituents and from that of the collocates stored in DEFIDIC. Consider the following WordNet-inspired attachments [collocates to the left of the arrow were found in the BNC, those to the right are DEFIDIC collocates they were related to]:

heavy *pause* ↔ heavy [menstrual] *period*
clear *head* ↔ clear *lead*
heavy *gas* ↔ heavy *attack*
heavy *pitch* ↔ heavy *sky*

WordNet's main drawback by far is its computational cost: searching through the WordNet taxonomy as described above takes 30 times as long as an MSS computation and a search through Roget's put together. This duration, however, could be cut considerably by reducing the scope of the search — mainly by reducing the hyper/hyponym search to two levels, and by dropping the search for a common hypernym.

Roget's can only be expected to provide lists of synonyms (be they sometimes very loose ones), which it does reliably enough. Roget's thesaurus (and especially our public domain version) shows its age by the many literary and Latin terms it contains, as opposed to its (comparatively) poor coverage of technical terms. Roget's tends to group words that are not usually associated any more, resulting in attachments the OH and CR lexicographers would never have thought of:

heavy *industry* ↔ heavy *work*
to cut *growth* ↔ to cut *grass*
to cut *speed* ↔ to cut *grass* (nb: this is drugs slang, not literary or old-fashioned)

On the other hand, Roget's is very good at identifying near collocates that are completely overlooked by the other two databases. In the following examples, attachments are due mainly or entirely to Roget's:

to clear a *passage* ↔ to clear a *road* or a *path*
a clear *perception* ↔ a clear *impression*
to cut a *cord* ↔ to cut a *rope*
heavy *footfalls* ↔ heavy *steps*
heavy *pressure* ↔ heavy *weight* or *load*
light *trance* ↔ light *sleep*

The strong point of metalinguistic slot sharing, as mentioned above, lies in its surprising ability to associate words that are related by none of the traditional semantic relationships. Consider the following examples, where MSS was mainly or solely responsible for the attachment:

a bright *idea* ↔ a bright *person*
a bright *window* ↔ a bright *room* (bright meaning 'not dark')
Bright Young *Things* of the shadow cabinet ↔ bright *person*
a heavy *garment* ↔ heavy *fabric*

a light *area* ↔ a light *room* or *house*
light *food* ↔ light *wine* or *meal*

MSS, because it relies on DEFIDIC's collocates, works best with simple, everyday words — lexicographers, as a rule, avoid including technical or otherwise complex terms in the metalinguistic apparatus. Since it works with words that were intended as thesauric heads, all 'marked' or specialized terms are absent from its base. This explains, for instance, that MSS could not match *schoolboy* and *lad* with *child* (even *boy/child* receives a score of only 4). As could be expected, MSS is actually not very good at matching synonyms: real synonyms, logically, should not co-occur within collocate lists. Its main drawback so far is the noise generated by the word *person*: while MSS tends to assign scores of 4 or 8 somewhat haphazardly (just as WordNet all-too easily finds common hypernyms), the problem is much more serious in the case of *person*, which appears 3250 times in our MSS database (and almost 8000 times in the collocate lists of the whole dictionary). This frequency is justified by the fact that *person* applies to all words potentially concerning human beings, but its consequence is that *person* matches just about anything with relatively high scores. A simple way to get rid of that noise would be to systematically halve all MSS scores involving *person*: experience has shown that such halving is mostly sufficient to let other, more correct matches predominate, and that it has little effect on the correct MSS scores involving *person* — these are so high anyway that halving them rarely takes away their top position.

5. Conclusion

Apart from shedding some distressing light on the limitations and shortcomings of three collocate-matching methods, this experiment has also shown how necessary they are: without them, the DEFI look-up program would have ended up with an uncertainty rate of 100% when confronted with all the collocates analyzed here. Further work will consist in improving the WordNet search routines, in reducing MSS noise and, if possible, in implementing a Roget's search with a more recent version of the thesaurus. Beyond that, any context-sensitive approach will have to take into account more than the immediate context.

6. Notes

1 For a more detailed description of the matcher see Michiels (1998).

2 Public domain Roget's thesaurus downloaded from the Project Gutenberg WWW site (www.gutenberg.net)

7. References

Bauer, D., Segond, Fr., Zaenen, A. (1995): *Locolex: the translation rolls off your tongue* (Technical report MLTT).

Corréard, M.-H., and Grundy, V., eds (1994): *Oxford-Hachette French Dictionary* (Oxford: O.U.P.).

Dufour, N. (1997): *Merging two DEFI dictionaries*, DEFI technical report, Liège, available from http://engdep1.philo.ulg.ac.be/michiels/defi.htm.

Dufour, N. (1997): DEFIDIC, a lexical database for computerized translation selection, in *RISSH* vol. 33, Liège.

Duval A. and Sinclair L.S., eds. (1993): *Collins-Robert French Dictionary* (Glasgow: HarperCollins).

Fontenelle Th. (1997a): Using a bilingual dictionary to create semantic networks, in *International Journal of Lexicography, X*, 4.

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

Michiels A. (1996): *The DEFI project: start here*. Technical report available on Defi's WWW site at http://engdep1.philo.ulg.ac.be/michiels/defi.htm.

Michiels, A. (1998): The DEFI matcher, in *EURALEX'98 Proceedings*, University of Liège.

Miller G. A., Beckwith R., Fellbaum Ch., Gross D. and Miller K. J. (1990): Introduction to WordNet: An On-line Lexical Database, in *International Journal of Lexicography, III*, 4, pp. 235-244.

Montemagni S., Federici S. and Pirrelli V. (1996): Example-based Word Sense Disambiguation: a Paradigm-driven Approach, in *EURALEX'96 Proceedings*, University of Göteborg, pp. 151-159.