Thierry Fontenelle, University of Liège

Collocation acquisition from a corpus or from a dictionary: a comparison

ABSTRACT: This paper focuses on the extraction of collocations from the "Collins-Robert English-French, French-English dictionary". The extraction programme, based on the WordCruncher Text Retrieval software package, is illustrated by the study of the combinatory properties of the word PRICE. The co-occurrence knowledge extracted from the dictionary is then compared with similar data retrieved from a statistically-processed corpus. The two techniques are assessed and shown to be complementary and mutually enriching.

1. Introduction

The development of natural-language processing systems requires access to large lexicons including morphological, semantic and syntactic information about thousands of lexical items. Some researchers are convinced that it is not desirable to code lexical items manually and that it is more efficient to start from already-existing resources, whether dictionaries (machine-readable dictionaries, lexical data bases) or large textual corpora (see i.a. Calzolari 1989; Byrd 1989, 67-79). However, some NLP applications, such as machine translation, need to combine lexical knowledge with knowledge of the world to ensure consistency in the disambiguation process. This requirement has led to the concept of Lexical Knowledge Bases (LKB) (Calzolari 1989). The distinction between lexical information and knowledge of the world is not always clear-cut and some types of information may be considered as borderline cases. This paper describes a technique for extracting co-occurrence knowledge, a type of lexical knowledge, from a machine-readable dictionary. This technique is compared with other methods which make use of statistical tools to process large corpora in order to acquire collocational information (Church & Hanks 1990, 22-29). The ultimate aim is to capture and formalize idiosyncratic collocation constraints that can as yet only be found in specialized printed dictionaries such as the "BBI Combinatory Dictionary of English" (Benson et al. 1986).

This paper shows that dictionary analysis and corpus analysis complement each other and enable researchers to produce better descriptions of lexical items. The word PRICE is taken as an example to illustrate the types of data that can be obtained when the two techniques are applied.

2. A machine-readable dictionary

The technique described below is based on the machine-readable version of the "Collins-Robert English-French French-English dictionary" (Atkins & Duval 1978). The magnetic tapes of the dictionary were made available to our department at the University of Lige for research purposes under contract with the publishers. In order to ensure quick access to the dictionary files, it was decided to use the WordCruncher Text Retrieval software package (formerly BYU Concordance). This software package runs under MS-DOS and some preparation was needed before the dictionary could be exploited.

WordCruncher has generated a general index (with the frequency of occurrence) of all the words that appear in one part of the dictionary. A "word" for WordCruncher is defined as any string of characters recorded according to the "character sequence file". The new version of the dictionary now distinguishes between the following:

1. Information which appears in italics in the printed version (mainly metalinguistic information – such as part of speech, subject field, and co-occurrence restrictions); such information now appears between <and> and is recorded as such in the general index compiled by the system.
2. Non-lexical information (such as punctuation marks and non- ASCII characters ...).
3. The English words (now in capital letters).
4. The French words (now in small letters).

The following example shows how words in italics, English words and French words are distinguished. The code {u89} represents the symbol * which is used in the printed version to indicate informal words or expressions.

> ROCKET 2 <vi> <[prices]> monter en flèche. <(fig)> TO ROCKET TO FAME devenir célèbre du jour au lendemain ; HE WENT ROCKETING {u89} PAST MY DOOR il est passé en trombe devant ma porte

3. Collocations: "Words shall be known by the company they keep"

The notion of COLLOCATION has proved influential in the design of learner's dictionaries (Mackin 1978, 149-165). The term "collocation" refers to the idiosyncratic syntagmatic combination of lexical items and is independent of word class or syntactic structure (compare the famous series of examples: to argue strongly – a strong argument – the strength of an argument). It is now a well-established fact that the learners of a language tend to memorize word associations. In many cases, they are even able to predict one element once they are given the other.

Collocation is a key concept not only in applied linguistics but also in NLP and, especially, in natural language generation and machine translation. It is in this latter field that word sense assignment is crucial for the selection of the appropriate word in the target language. The disambiguation process often relies, among many other factors, on co-occurrence knowledge and the main problem is twofold:

a. Collocation constraints being idiosyncratic, they need to be formalized to be put to good use in an NLP system and to avoid such oddities as * "putrid butter" instead of "rancid butter". (Nirenburg 1989, 43-65).

b. A methodology must be developed to acquire this knowledge on a large scale and, possibly, automatically, whether from corpora (Church & Hanks 1990, 22-29) or from already-existing machine-readable dictionaries (Boguraev 1991, 227-260).

4. Collocations in the Collins-Robert dictionary

The dictionary provides a wealth of information on co-occurrence knowledge and collocations in the form of typical objects, subjects or noun complements of adjectives. A systematic approach has been adopted by the lexicographers to consistently code collocational information in italics:

- square brackets, [], indicate a typical noun subject of the headword, with verb entries;
- the absence of parentheses indicates a typical noun object of a transitive verb or a typical noun that can be modified by an adjective

The following examples illustrate this approach:

ABOLISH *vt law* abolir, abroger TRUMPET *vi [elephant]* barrir
REPEAL *vt law* abroger, annuler ADDLED *adj egg* pourri

It should be noted that square brackets can also be used within noun entries to refer to a typical noun complement of the headword, for example: LEAF 1 *n* (a) *[tree, plant]* feuille.

As can be seen from the above examples, adjective-noun and verb-noun collocations will be found under the adjective and the verb respectively. Using Hausmann's terminology (Hausmann 1985, 118-129), we say that the BASE of the collocation is the element in italics whereas the COLLOCATOR is the headword. Such practice is commonly found in general-purpose dictionaries.

The WordCruncher organisation of the Collins-Robert dictionary makes it possible to instantaneously retrieve all the occurrences of a given word in italics, together with the headword under which this item is to be found. The intermediate file produced by WordCruncher can then be submitted to a simple AWK programme which automatically assigns the syntactic/semantic link between the word in italics (the base) and the headword (the collocator) on the basis of part-of-speech information combined with typographical information, such as square brackets and parentheses. The four links that are assigned by our programme are: adjective; modifier_of_noun; object_of; subject_of.

An experiment has been carried out with the word PRICE. A sample output of the programme applied to PRICE runs as follows: query on "price" or "prices" in italics extracted from the dictionary; the complete list comprises more than 200 items.

adjective PROHIBITIVE	object_of FIX
subject_of COLLAPSE	modifier_of_noun LOWNESS

The following adjectives can modify PRICE (adjective):

ATTRACTIVE, AVERAGE, COMPETITIVE, CURRENT, DEAR, DECREASING, DIMINISHING, EXCESSIVE, EXORBITANT, EXTORTIONATE, EXTRAVAGANT, FORWARD, GIVEAWAY, GOING, HEFTY, HIGH, INFLATED, KEEN, LOW, MEAN, MINIMUM, MODERATE, MODEST, NET, OPENING, OUTRAGEOUS, OUTSIDE, PRETTY, PROHIBITIVE, REASONABLE, REGULAR,

RULING, SACRIFICIAL, SET, SHOCKING, SOARING, SPECIAL, STABLE, STEADY, STEEP, STIFF, UNBEATEN, UNREASONABLE, USUAL, WHOLESALE

The following transitive verbs take PRICE as object (object_of):

ADVANCE, AGREE, ASK, BEAT DOWN, BOOST, BRING DOWN, BUMP UP, CONTROL, DEPRESS, DETERMINE, DOUBLE, DROP, ENHANCE, ESCALATE, ESTIMATE, EVEN OUT, FIX, FREEZE, INCREASE, INFLATE, JACK UP, KNOCK DOWN, LAY DOWN, LOWER, MARK, MARK DOWN, MARK UP, NAME, PEG, POLICE, PUSH UP, PUT UP, QUOTE, RAISE, REALIZE, REDUCE, ROUND DOWN, ROUND UP, SEND DOWN, SLASH, STIPULATE, TAKE OFF, UP

The following intransitive verbs take PRICE as subject (subject_of):

ADVANCE, BE ON THE UPGRADE, BE UNDERSTOOD, BE UP, BOOM, CLIMB STEEPLY, COLLAPSE, COME DOWN, COME INTO FORCE, DECLINE, DECREASE, DIP, DOUBLE, DROP, EVEN OUT, FALL, FLUCTUATE, GET DEARER, GO DOWN, GO THROUGH THE ROOF, GO UP, HARDEN, HIKE, HIT THE CEILING, HIT THE ROOF, INCREASE, JUMP, KEEP UP, LEAP UP, LEVEL OFF, LEVEL OUT, LOWER, MOUNT, PICK UP, PLUMMET, PLUNGE, RECEDE, RISE, RISE STEEPLY, ROCKET, SAG, SHOOT UP, SINK, SKYROCKET, SLUMP, SOAR, SPIRAL, SPIRAL UP, STAND AT, STEADY, TAKE A PLUNGE, TOBOGGAN, WEAKEN

Acquiring collocational information about PRICE from the Collins-Robert and studying the combinatory properties of this item enables us to answer the following three questions:

(1) What can prices be like? (cf. adjective)
(2) What can prices do? (cf. subject_of)
(3) What can be done to prices? (cf. object_of)

Answering these questions is undoubtedly crucial in the dictionary-making process. Having access to such data enables the lexicographer to specify and formalize the environment in which it is likely to occur. Here, the set of verbs that can collocate with PRICE can be divided into three further sub-classes:

a) verbs referring to a rise in prices – synonyms of INCREASE
b) verbs referring to a fall in prices – synonyms of DECREASE
c) verbs referring to the stability of prices

At this juncture, it should be stressed that the use of PRICE in italics covers two different concepts. First of all, it refers to the actual lexical item PRICE that can collocate with all the words that are given above. But it also refers to the head of a thesauric class to which words such as SHARE, STOCK, DOLLAR, POUND, and all other currencies belong. Caution should therefore be exercised in analysing words in italics since no dictionary makes a distinction between collocations that are restricted to a single item and collocations with a semi-restricted set of items. When we say that THE FIRM'S SHARES JUMPED TO £12, we mean that the PRICE of the firm's shares jumped to £12. This type of metonymy should of course be taken into account and such semantic extensions should be captured in a thesaurus-based NLP system (Michiels & Noël 1982, 227-232).

5. Ergativity

The method of locating collocations suggested in this paper makes it possible to discover lexical items that display particular syntactic properties. A closer look at the list of verbs that collocate with PRICE shows that some verbs appear both in the "object_of" class (i.e. transitive verbs that can have PRICE as typical object) and in the "subject_of" class (i.e. intransitive verbs that can have PRICE as typical subject). These verbs are: ADVANCE, DOUBLE, DROP, EVEN OUT, INCREASE, and LOWER. They display the so-called causative/inchoative alternation (Atkins et al. 1988, 84-126), which means that they can be both transitive and intransitive and that the object of the transitive construction can be used as the subject of the intransitive verb. In terms of semantic roles, these "ergative" verbs involve an AGENT argument and a PATIENT argument. The patient is the subject when there is no explicit agent, as is seen in the intransitive construction. In the present case, PRICE is the patient argument that undergoes a change of state. This property accounts for the following alternation:

> IT IS UNLIKELY THAT PRICES WILL DROP ANY FURTHER vs THEY ASKED US TO DROP OUR PRICES BY FIVE PER CENT

Since the number and the nature of semantic roles must be identified and assigned at definition level, this method seems adequate to automatically retrieve and code ergative verbs. Compare this with the method for extracting ergative verbs from the Longman Dictionary of Contemporary English described in Fontenelle & Vanandroye (1989, 11-39) or in Boguraev (1991, 227-260).

6. Corpus analysis of PRICE

A dictionary such as the Collins-Robert typically aims for BREADTH of description. Unlike MRDs, however, a corpus is more useful to analyse frequent phenomena in DEPTH. One of the major problems with dictionaries is that the collocations they include are often arbitrary and that we have no evidence that they reflect the actual behaviour of words. Only a corpus can provide us with statistical information on the frequency of combinations. The word PRICE was therefore examined with a view to describing its environment in actual texts. Concordances for PRICE/PRICES were extracted from the Oxford University Press Pilot Corpus and scrutinized individually[1]. A database was created to record the following types of information for each concordance:

- adjectives that modify PRICE,
- transitive verbs that take PRICE as object,
- verbs that take PRICE as subject,
- nouns which pre-modify PRICE,
- nouns which PRICE pre-modifies,
- prepositions associated with PRICE.

The analysis of this data base made it possible to discover which adjectives, nouns, verbs, and so on, occurred most frequently in combination with PRICE. The two sets of data - the dictionary data and the corpus data – were then compared and examined. Predictably enough, the corpus-based approach yielded many more possible combinations than

the Collins-Robert, and the dictionary mentioned information which, in most cases, was to be found in the corpus but was not necessarily typical or central. The analysis of the corpus revealed the following facts:

1. The ten most frequent adjectives that modify PRICE are – in descending order : high (77 occurrences), higher (53), low (42), lower (37), reasonable (33), average (30), asking (29), rising (21), reduced (21), latest (21).

2. The eleven most frequent verbs that take PRICE as subject are in descending order: be (240), rise (83), fall (63), go up (40), range (17), include (16), soar (15), drop (15), tumble (13), reflect (11), come down (11).

3. The ten most frequent transitive verbs that take PRICE as object are: pay (150), raise (29), set (22), offer (22), increase (21), cut (20), charge (20), reduce (17), push up (15), put (13).

4. The most frequent compounds with PRICE modifying a noun are: p.rises (60), p.index (42), p.increase (29), p.tag (26), p.rise (17), p.range (12), p.panel (12), p.inflation (12), p.freeze (11), p.control (11).

5. The noun PRICE is most often preceded by the following nouns: share (174), house (113), oil (57), retail (47), purchase (30), market (29), petrol (25), property (22). It will be noted that these nouns do not always refer to goods that can be sold and have a price – such as oil, petrol, houses, shares or property: the purchase price is not the price of a purchase and the market price is not the price of a market. Such fixed combinations should therefore be (and are) entered separately in the dictionary.

6. In 267 instances, PRICE is preceded by AT (e.g. sold at a reasonable price)
In 366 instances, it is followed by OF (e.g. the price of oil, of coffee)
In 178 instances, it is followed by FOR (e.g. he paid the price for the software)

Church & Hanks (1990, 22-29) have shown that it is possible to estimate word association norms from large tagged corpora. Their statistical calculation reveals that the patterns discovered in the association ratio tables can help lexicographers organize a concordance in detecting significant collocations. A comparison with the list of collocations drawn from the corpus and selected by Ken Church's statistical routines shows that the twelve most frequent collocates of PRICE are: SHARE, RISES, FOR, PURCHASE, PAID, RETAIL, PAY, ITS, OIL, INDEX, ASKING, AT. This is very interesting since it also gives access to the most typical prepositions associated with the noun (FOR, AT) or to the most typical pre-modifiers (SHARE, OIL), a type of data about which the dictionary is rarely informative – except in examples. It should however be noted that the combination of PRICE with share/oil/house/etc is transparently analyzable and probably depends on the type of texts in the corpus. The inclusion of such information in a dictionary would therefore be highly questionable. One of the other drawbacks is that this routine does not say anything about the link between the collocates: it does not indicate, for example, whether RISES is a plural noun or a third person singular verb.

It is fairly easy to understand why the Collins-Robert dictionary does not record some very frequent and significant collocations. A GOOD PRICE occurs 12 times in the corpus but no information is given under GOOD in the dictionary because this adjective can collocate with practically any noun. It is therefore hard to formalize its range. It will then be mentioned under PRICE (HE GOT A GOOD PRICE FOR IT). The same is true of other elements which have a wide collocational range such as ANY (in AT ANY PRICE) or the

verb BE (THE PRICE IS £5). Verbs such as RANGE, INCLUDE, TUMBLE, PAY, SET, OFFER and many others should, however, have contained some reference to their collocability with PRICE in the dictionary. Moreover, many collocators found in the dictionary, such as EXTORTIONATE, ATTRACTIVE, PROHIBITIVE, or FLUCTUATE, occur only once in the corpus, which is statistically insignificant. This clearly demonstrates that intuition may not reflect typical and central usages as evidenced by a well-balanced and carefully designed corpus.

Interestingly, the Collins-Robert lexicographers have included collocations which, intuitively, seemed to be frequent but which do not appear in the corpus. These include GOING PRICE, SACRIFICIAL PRICES, RULING PRICE, DEAR PRICE, MEAN PRICE, and UNREASONABLE PRICE. According to the dictionary, we can also ADVANCE prices, we can BEAT them DOWN, ENHANCE them, ESCALATE them, EVEN them OUT, LAY them DOWN, MARK them UP, POLICE them, ROUND them DOWN or UP, SEND them DOWN, STIPULATE them, TAKE them OFF or UP them but the corpus does not give evidence that these are typical, central or even possible combinations. The Collins-Robert also tells us that prices can ADVANCE, DECREASE, EVEN OUT, HARDEN, HIKE, SAG, SKYROCKET, HIT THE ROOF, TAKE A PLUNGE or TOBOGGAN but these usages are not attested in the corpus either. We might of course wonder whether an 11-million-word corpus is large enough to capture such combinations. The size and the scope of the corpus are obviously key factors and lexicographers are aware that they should be taken into account, witness the British National Corpus initiative which aims at creating a corpus of 100 million words of contemporary spoken and written British English.

7. Conclusions

It is a well-established fact that designing a lexicon for NLP systems requires large lexical knowledge bases. The purpose of this paper was to show that computational dictionary analysis and corpus analysis are complementary and that they may be put to good use on a lexicographer's workbench since fast access to large bodies of lexical material is essential in order to compile dictionaries for human beings or for machines. I have described a technique that identifies and retrieves collocations from the Collins-Robert dictionary and assigns a syntactic tag that represents the surface link between the base and the collocator. It proves to be efficient mainly with respect to adjective-noun and verb-noun collocations, but may fail to give access to central and representative linguistic usages. The description of lexical items can therefore be enriched by lexical data extracted from very large corpora submitted to statistical analyses à la Church and Hanks which yield more information with respect to accompanying prepositions, noun-noun combinations or typical usages. The latter approach, however, also yields transparently analyzable combinations and depends heavily on the size and the design of the corpus. The two methods being complementary, they should therefore be combined synergically to construct NLP dictionaries.

Endnotes

1 The 3,715 citations for the word-forms PRICE/PRICES are drawn from the 11.1 million word Oxford University Press Pilot Corpus (situation in January 1991). I wish to thank Sue Atkins and Jeremy Clear from OUP through whose courtesy these concordances and the collocations selected by Ken Church's statistical routines were made available to me. My thanks also go to the anonymous referees whose valuable comments helped me improve an earlier draft of this paper.

Bibliography

ATKINS, B.T., DUVAL, A. (1978): Collins Robert French-English English-French Dictionary. Collins/ Dictionnaires Le Robert, Glasgow/Paris.

ATKINS, B.T., KEGL, J., LEVIN, B. (1987): "Anatomy of a Verb Entry: from Linguistic Theory to Lexicographic Practice". In: International Journal of Lexicography 1:2. O.U.P. Oxford.

BENSON, M., BENSON, E., ILSON, R. (1986): The BBI Combinatory Dictionary of English. John Benjamins Publishing Company. Amsterdam-Philadelphia.

BOGURAEV, B. (1991): "Building a Lexicon: The Contribution of Computers". In: International Journal of Lexicography. 4:3. O.U.P. Oxford.

BYRD, R. (1989): "Discovering Relationships among Word Senses". In: Dictionaries in the Electronic Age: Proceedings of the Fifth Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary. Oxford.

CALZOLARI, N. (1989): "Lexical Databases and Textual Corpora: perspectives of integration for a Lexical Knowledge Base". In: Proceedings of the First International Lexical Acquisition Workshop. Ed. by U. Zernik. Detroit.

CHURCH, K., HANKS, P. (1990): "Word Association Norms, Mutual Information and Lexicography". In: Computational Linguistics. 16:3.

FONTENELLE, T., VANANDROYE, J. (1989): "Retrieving Ergative Verbs from a Lexical Data Base". In: Dictionaries. Vol. 11. Dictionary Society of North America.

HAUSMANN, F. (1985): "Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des Lexikographischen Beispiels". In: Lexikographie und Grammatik. Ed. by Bergenholtz & Mugdan. Niemeyer. Tübingen.

MACKIN, R. (1978): "On Collocations: 'Words shall be known by the company they keep'". In: In Memory of A.S. Hornby. Ed. by P. Strevens. Oxford University Press. Oxford.

MICHIELS, A., NOËL, J. (1982): "Approaches to Thesaurus Production". In: Proceedings of COLING82, North-Holland. Amsterdam.

NIRENBURG, S. (1989): "Lexicons for Computer Programs and Lexicons for People". In: Dictionaries in the Electronic Age: Proceedings of the Fifth Annual Conference of the University of Waterloo Centre for the New Oxford English Dictionary. Oxford.

WORDCRUNCHER Text Retrieval Software. Brigham Young University. Provo, Utah.

KEYWORDS: computational lexicography, collocation, lexical acquisition, bilingual dictionaries, corpus analysis