COLLOCATION ANALYSIS OF A SAMPLE CORPUS USING SOME STATISTICAL MEASURES: AN EMPIRICAL APPROACH.

ANTONIA SÁNCHEZ MARTÍNEZ
Escuela Oficial de Idiomas, Murcia

ABSTRACT. *The object of this research is to produce a comparative study of the quantitative methods most commonly used for collocation extraction. These are: T-score, Z-score and Mutual Information. In order to make this comparison, we will take into account the different existing definitions for collocation. We will conclude that depending on the definition we consider, the researcher will use one statistical method or another. We want to highlight the importance of the statistical method employed; it is essential to use the most effective method when extracting collocational information from a corpus, since the reliability of the results will depend on the effectiveness of the method.*

RESUMEN. *El objetivo de esta investigación es hacer un estudio comparativo de los métodos cuantitativos de extracción de colocaciones más comúnmente utilizados. Éstos son el Coeficiente T, el Coeficiente Z y la Información Mutua. Para realizar esta comparación tendremos en cuenta las distintas definiciones de colocación, y así concluiremos que dependiendo de la definición de colocación por la que se opte, el investigador usará un método estadístico u otro. Queremos poner de relieve la importancia del método estadístico usado; es fundamental utilizar el método más efectivo a la hora de extraer la información colocacional de un corpus; de la eficacia del método usado dependerá la fiabilidad de los resultados.*

1. DEFINITION OF COLLOCATION

We will regard collocations as the statistically significant co-occurrence of words within a short span in a text. A collocation of two of more words may be simply due to grammatical rules, such as *determiner + noun* in the case of *the cat*; or it might be the case that those collocates co-occur together because they are part of an idiom, saying or proverb. But our interest moves beyond simple co-occurrence to significant collocations with a degree of signification that can be statistically measured. In a statistically significant collocation, one lexical item requires the presence of another or others. These significant collocations, which can neither be explained in terms of syntax nor be considered as a semantic unit (idiom), attract our attention in this study.

Keeping in mind the nature of the researcher's study, he will have to decide:

(1) The number of collocates subject to study on each side of the node word. The number of words within the span varies from Clear's (1993) window of 4 to Martin's window of 10 (Martin et al 1983).

(2) Whether grammatical words are included within the span or not. Investigating phrasal verbs or the prepositions adjectives tend to be followed by, would obviously require the presence of those grammatical words.

(3) Whether the linguistic corpus is going to be pre-processed or not. As Stubbs (1996: 172) observed "different word forms can have quite different collocates". One surprising example is the one he gives. The word *educate*: while *education* collocates with words referring to institutions (higher, secondary, university), the term *educate* co-occurs with its synonyms *enlighten, help, inform, train*. Also, the form *educated* is repeatedly followed by *at*.

2. COLLOCATIONAL ANALYSIS

Our aim is to identify the collocations in a sample corpus and to determine their significance using the *t-score*, *z-score* and *mutual information* values:

- The *z-score* compares the observed frequency between a node and its collocates to the expected one and evaluates the difference between these values by means of a standard deviation.
- The *mutual information* "*I(x,y)*, compares the probability of observing a word *x* and a word *y* together (the joint probability) with the probabilities of observing *x* and *y* independently (chance)" Church et at (1991: 120).
- The *t-score*, though very similar to the *z-score* formula, takes frequency into account, which is said to provide more accuracy dealing with those words with a relatively low frequency.

Since those are the three significance measures most commonly used in the extraction of statistically relevant collocations (Stubbs 1996, Church and Hanks 1990, Clear 1993, Barnbrook 1996), we will use them in our study comparing their results in order to highlight their respective advantages and disadvantages.

3. CORPUS SELECTED

The corpora we have used are Corpus Collections A and B, both published by Oxford University Press. Each corpus consists of approximately one million words, totalling of 2.047.903 words, containing written and spoken language samples of various domains. Our node word, *time*, has been selected among those words with a significant number of frequencies within the corpus.

4. NODE ANALYSIS

The 3.372 occurrences of *time* are taken together with a span of five words on each side of the node word. The result is a context sample text (Table 1) that we can see in a KWIC concordance list.

```
    this year 1364." Although time seems always to have been
 a relatively short period of time, since the illness is beyond
       a little for the first time but now it has stopped.
        a careful note of the time of full eclipse, which was
             a few lines at a time. In June, his botany lecturer
  a good one: without change time could not be recognized, whereas
```

Table 1. Sample Concordance list for the node word *time*.

Examining in more detail the co-occurring words within the span, we get the following figures as a result: there is a total of 712 different collocates of the word *time* in the contextual sample taken; the article *the* is the one with a higher number of co-occurrences with a total of 2742. Here, we have included a short list of the most common collocates (Table 2) of the node word together with their number of co-occurrences.

WORD	TOTAL		
THE	2742	BUT	208
AND	714	SAME	204
FOR	646	FROM	189
THAT	399	HIS	184
WAS	391	WITH	179
THIS	355	WHEN	160
FIRST	258	NOT	150

Table 2. List of co-occurring words with *time* ordered by frequency.

5. Results

We will proceed to compare the results of the three significant measures mentioned above: *t-score*, *z-score* and *mutual information (MI)*. In order to do that, a table is presented where all collocates are ordered depending on their respective values for each method. (Table 3)

Z-SCORE	MI SCORE	T-SCORE
SAME	ZURVAN	FOR
HAS	RIPE	FIRST
SPACE	CUES	SAME
CUES	CYCLICAL	HAS
ABOUT	BEFORE	THIS
FIRST	SPENDS	SPACE
BEFORE	SPEND	ABOUT
MUCH	MUCH	THE
SPENT	SECOND	WHEN
SECOND	SPACE	LONG
COULD	FINITE	WAS
FOR	LAPSE	MUCH
SPEND	WASTED	SOME
INTO	ABOUT	BEFORE
ZURVAN	KERR	COULD
RIPE	SPENT	CUES
LONG	MEASURING	INTO
THIS	HAS	SPENT
FULL	OCCURRING	FULL
WHEN	WASTE	HALF
USE	ETERNAL	SECOND
SUCH	USE	SINCE
PARTY	COULD	COME
WASTE	CUE	SUCH
CYCLICAL	INTO	SPEND

Table 3. *Significance measures compared. In red: words occurring in the three columns, in blue words occurring in z-score and MI columns and in green words occurring in z-score and t-score columns.*

6. Conclusions

There are two major types of differences shown here: words which are included in one list but not in another and words whose ranks differ between the lists. The first noticeable thing is that 44% of all the words in the table occur in the three lists. Apart from those, none of the words left occurring in the *MI* list can be found in the *t-score* list.

There are 5 collocates, 20% of the words in the table, in the *z-score* and *MI* columns which cannot be found in the *t-score* results, namely: *cyclical, ripe, use, waste* and *zurvan*. Most of them have low frequencies in the whole corpus, especially in the cases of *cyclical, ripe* and *zurvan*. On the other hand, there are a number of words which can be found in the *t-score* column and not in the other two, they are: *come, half, since, some, the* and *was*; these 24% of collocates have indeed a very high frequency in our corpus. While *MI* and *z-score* results are closer, differences between the *MI* and the *t-score* lists are even bigger since fourteen of the collocates in the table are different; that is, 64% of the collocates in the *MI* and *z-score* columns are the same while the percentage of similarity between MI and t-score columns is just 44%.

Having a closer look at the significant collocates in the *z-score* and the *MI* tables, we find lexical items with a relatively low total frequency in the corpus; some of which, *ripe* or *zurvan,* are themselves very rare occurrences in the corpus. Although they appear in the top positions within the *MI* list, we cannot be sure whether a collocation that has been observed only a few times (5 in the case of *zurvan* and 8 for *ripe*) is really reliable and can be taken as a guide to co-occurrence patterns. As we said before, the majority of words in the *z-score* and the *MI* lists have, to some extend, low frequency values in the whole corpus; at least, much lower than the collocates the *t-score* list has, among which we can find the definite article *the* with the highest occurrence within the corpus.

The significance of low frequency co-occurring words is artificially inflated by both the *MI* and *z-score* measures; being the *t-score* a much more reliable measure when dealing with collocates that have a low total frequency in corpus. At the same time the *MI* and *z-score* downgrade collocates with a relatively high frequency in the whole text, while the *t-score* gives prominence to very frequent words in the corpus.

In general, the *t-score* will be a lot more likely to highlight frequently recurring items (many of which will be grammatical words such as prepositions, personal pronouns, determiners and particles) together with fully lexical words strongly associated with the node word. On the other hand, the *MI* gives prominence to technical phrases, idioms, proverbs and fixed compounds. Important differences can be found between the three statistics compared and the information they provide; though the *z-score* and *MI* values are a lot more similar than the *t-score*.

It is very difficult to determine which one would be the ideal measure for collocation analysis; we believe the researcher should take advantage of the different perspectives provided by the use of more than one measure and therefore use as much information as possible in exploring collocations. Factors such as the purpose of the research should be taken into account, and the lexicographer should decide which statistic is the most appropriate for his study depending on his definition of collocation.

As explained above, collocations can be defined as the probability of two items co-occurring together within a determined span being greater than might be expected from pure chance. This definition is the one taken by most authors within the Firthian tradition: Greenbaum (1974), Geffroy et al (1973), Berry-Rogghe (1973), Nattinger (1980), Cowie (1981), Martin et al (1983), Cruse (1986), Church and Hanks (1990), Benson (1995), Mel'čuk (1998), Smadja (1993) and Clear (1993). In this sense, what matters are statistically significant collocations, which are the ones "in which the two items co-occur more often than could be predicted on the basis of their respective frequencies and the length of the text under consideration" Martin et al (1983: 84). It is really difficult to state which, the *MI* and *z-score* on the one hand, or the *t-score* on the other, would be the best measure to be adopted by all these authors. Perhaps the *t-score* would be the most appropriated one, though the researcher could also benefit from the *z-score* and *MI* values, taking into account that they can be misleading when dealing with very low corpus frequency collocates.

However, some other authors, namely Kjellmer (1987) and Kita et al (1994) do not make any difference between collocations, idioms or compounds and regard the former as the co-occurrence of two lexical items within a specified co-text. For these two authors, who consider collocations differently, the use of the *MI* and *z-score* value would be preferred since those calculations focus their attention on the more idiosyncratic collocates of the node: technical terms, fixed phrases and compounds.

The main benefit of the techniques analysed is that they focus our attention on the company words keep (Firth 1968). A corpus-based analysis of word co-occurrence yields a wealth of data for the benefit of lexicographers; quantitative analysis of the occurrence of word forms allows us to extract the significant collocates within a corpus, so that patterns about the behaviour of words can be assessed.

Those significant collocates extracted provide us with very useful information that can be used in second language teaching and learning, sentence generation, information retrieval, machine translation and the compilation of dictionaries, helping the lexicographer to sharpen his definitions. Hence, collocational information is a very valuable one and although many analysts have devised quantitative statistical methods for its extraction, these methods also present some problems, namely:

- Determining an optimal span
- Establishing a cut-off point for significance for each of the statistical measures
- Dealing with collocates with a low frequency in the whole corpus

The main conclusion after the comparison of three of the most widely used measures was that, though similar, their results differed especially between the *t-score* and the other two rather than between the *z-score* and the *MI* themselves.

Z-score and *MI* artificially inflate the values for those occurrences with a low total frequency in the corpus, while at the same time, downgrade those collocates with a high corpus frequency. Both these measures give high results to words that are neither lexically nor semantically related to the node word; common sense tells us that those two collocates are not really associated to our node word, but they co-occur by mere chance. Words with a low frequency in the corpus should be disregarded in the analysis when using either the *MI* formula or the *z-score*. Another option would be the use of the *t-score* formula when calculating the significance of these low corpus frequency words. It is very difficult to determine the best measure. There are important differences between the information provided by each method and the researcher should use all of them and, therefore, benefit from all the information the use of more than one measure provides.

REFERENCES

Barnbrook, G. (1996) *Language and Computers*. Edinburgh University Press.

Benson, M. (1995) "Collocations and Idioms" in R. Ilson (ed.) *Dictionaries, Lexicography and Language Learning*. CELT Documents 120; Oxford: Pergamon. 61-68.

Berry-Rogghe, G.L.M. (1973) "The computation of collocations and their relevance in lexical studies". In A.J. Aitkien, R. Bailey and N. Hamilton-Smith (eds) *The Computer and Literary Studies. Edinburgh*: Edinburgh University Press.

Church, K.W. and Hanks, P. (1990) "Word Association Norms, Mutual Information and Lexicography". *Computational Linguistics*, 16/1: 22-29.

Church, K.W., Gale, W., Hanks, P. and Hindle, D. (1991) "Using Statistics in Lexical Analysis". In U. Zernik (ed) *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*. Hillsdale, NJ: Lawrence Erlbaum Associates. 115-164.

Clear, J. (1993) "From Firth Principles: Computational Tools for the Study of Collocation". In M. Baker et al (eds) *Text and Technology*. Amsterdam: Benjamins. 271-292.

Cowie, A.P. (1981) "The Treatment of Collocations and Idioms in Learners's Dictionaries". *Applied Linguistics*, 2: 223-235.

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

Firth, J. (1968) “A synopsis of linguistic theory 1930-1955”. *In Selected Papers of J.R. Firth 1952-59*. Edited by F.R. Palmer. Bloomington, Indiana. Indiana University Press, 1-32.

Geffroy, A., Lafon, P., Seidel, G. and Tournier, M. (1973) "Lexicometric Analysis of Co-occurrences". In A.J. Aitkien, R. Bailey and N. Hamilton-Smith (eds) *The Computer and Literary Studies*. Edinburgh: Edinburgh University Press.

Greenbaum, S. (1974) "Some Verb-Intensifier Collocations in American and British English". *American Speech*, 49: 79-89.

Kjellmer, G. (1987) DECIDE Project. Web reference: http://engdep1.philo.ulg.ac.be/decide/

Kita, K., Kato, Y., Omoto, T., Yano, Y. (1994) “A Comparative Study of Automatic Extraction of Collocations from Corpora: Mutual Information vs. Cost Criteria”. *Journal of Natural Language Processing*, 1/1: 21-33.

Martin, W., Al, B. and van Sterkenburg, P. (1983) "On the Processing of a Text Corpus: From Textual Data to Lexicographical Information". In R. Hartmann (ed) *Lexicography: Principles and Practice*. London: Academic Press.

Mel’čuk, I. (1998) “Collocations and Lexical Functions”. In A.P Cowie (ed.) *Phraseology. Theory, Analysis and Applications.* Clarendon Press, Oxford.

Nattinger, J. (1980) "A Lexical Phrase Grammar for ESL". *TESOL Quarterly*, 14: 337-344.

Smadja, F. (1993) “Retrieving collocations from text: Xtract”. *Computational Linguistics*, 19(1): 143-177.

Stubbs, Michael. (1996) *Text and Corpus Analysis*. Blackwell Ed. Oxford.