D.E.L.T.A., Vol. 15, Nº 1, 1999 (141-149)

Word Sets, Keywords and Text Contents: An Investigation of Text Topic on the Computer*

(Iniciando a Lingüística do Corpus do Português: Explorando um Corpus para Ensinar Português como Língua Estrangeira)

A. P. Berber Sardinha *(Pontifícia Universidade Católica de São Paulo)*

Abstract: This study presents a methodology for the identification of coherent word sets. Eight sets were initially identified and further grouped into two main sets: a 'company' set and a 'non-company' set. These two sets shared very few collocates, and therefore they seemed to represent distinct topics. The positions of the words in the 'company' and 'non-company' sets across the text were computed. The results indicated that the 'non-company' sets referred to 'company' implicitly. Finally, the key words were compared to an automatic abridgment of the text which revealed that nearly all key words were present in the ahridgment. This was interpreted as suggesting that the key words may indeed represent the main contents of the text.

Resumo: Este estudo apresenta uma metodologia para a identificação de conjuntos de palavras coerentes. Oito conjuntos foram identificados inicialmente e posteriormente agrupados em dois conjuntos principais: um conjunto denominado 'companhia' e outro denominado 'não-companhia'. Estes dois conjuntos partilham alguns colocados, e portanto parecem representar tópicos distintos. A posição das palavras de ambos os conjuntos foi computada ao longo do texto analisado. Os resultados indicaram que os conjuntos 'não-companhia' se referiam indiretamente à companhia. Por fim, as palavras-chave dos conjuntos foram comparadas a um resumo do texto automático gerado por computador o qual revelou que quase todas as palavras-chave estavam presentes no resumo. Este fato foi interpretado como indício de que as palavras-chave representam o conteúdo central do texto.

Key Words: Corpus Linguistics; Teaching Portuguese as Foreign Language; Corpus-based description of Portuguese.

Palavras-Chave: Lingüística do Corpus; Ensino de Português como Língua Estrangeira: Descrição do Português baseada no Corpus.

* An earlier version of this paper was presented at the 7th International Systemic-Functional Workshop, Valencia, Spain, on 27 July 1995 under the title 'Segmentation and choice in written bussiness English', and published as DIRECT Working Paper 25.

0. Introduction

The aim of this paper is to propose a methodology for extracting coherent word sets from text. The relevance of the investigation lies in the fact that it can lead to the automatic identification of themes discussed in the text. Word set coherence is conditioned by two criteria: one, each *individual* set must sound coherent to the reader of the text; and two, the *set of sets* must present itself as coherent, that is, the individual sets must be related to each other by sharing the similar content. Word sets can be viewed as indicators of how the text can be partitioned into coherent segments.

The analysis is meant to be entirely inductive, that is, all of the interpretative categories will result from the analysis rather than be imposed onto it. Other studies have applied a bottom-up methodologies in the investigation of texts (Miall, 1992; Phillips, 1988; Scott, 1995; Wilson, 1993).

It is possible to derive word sets by studying word distribution statistically (e.g. Knott and Dale, 1993). A disadvantage of a pure statistical approach is that it may sacrifice validity. In our case, this means that the interpretation of the results would become too complex for the vast majority of researchers in applied linguistics.

In Systemic Linguistics there have been attempts at extracting word sets with only moderate use of statistics or none at all. For instance, Halliday (1992) outlines the possible lexical sets in a commercial letter. Similarly, Benson and Greaves (1992) describe and discuss the word sets in a letter in terms of the systemic category of *field*. Since they succeed in carrying out their analysis by computer, their study will serve as a basis for the present investigation. Their approach was based upon the intuitive identification of what was in their opinion the main clause or 'pitch clause' of the letter. In this clause, they decided that the word 'please' was central to it and extracted its collocates. The collocates were distributed into meaning sets. There are a few points in their technique which can be improved. As they suggest, comparisons of word frequency were carried out by using the absolute frequencies of the words in question, rather than frequencies relative to text size. Also, the key word in the text was chosen prior to the extraction of the field sets.

The technique proposed here uses the concept of 'key word' as implemented in the KeyWords tool of the WordSmith suite (Scott, 1995). Firstly, the text will be compared to a corpus of similar texts and the chi-square of

the distribution of each word type in the text and in the corpus will be computed. Secondly, the words will be sorted according to the significance of the chi-square statistic. At this stage there are two possibilities. Either the word is significantly more frequent in the corpus or it is significantly more frequent in the text. In the former case, the word will not interest us, since its frequency will be less than expected in the text and therefore will not characteristic of the text. In the latter case, however, the word will have been used in the text more frequently than expected and therefore it can be considered to characterize the text. Words which characterize the text by being unusually frequent in this way will be called KEY WORDS[1], although in the original proposal in Scott (1995) these would be called 'positive' key words. Finally, the collocates for each key word will be computed. Each set will then be formed by key words and their collocates.

The methodology described above can perhaps improve a few aspects of the approach applied by Benson and Greaves (1992). First, relative frequencies are compared by taking into account the difference between their expected and observed values (chi-square). Second, key words are not chosen intuitively, but by computing statistical significance.

The problem with relying on chi-square significance to identify key words is that high frequency words will often turn out to be key. But since high frequency words are in their majority function words, these words will not be key because function words are excluded from the analysis. Initially, all content words (nouns, main verbs, adjectives, adverbs) were allowed to become key. This was established following Eggins (1994), according to whom changes in field have 'an immediate impact on the content words used' (:68). However, it was later found that a further filtering was necessary so as to make sets as coherent as possible. Here we followed Halliday (1992), who in his analysis for lexical sets, included only nouns.

2. Collocates

Each word set is constituted by at least one key word and its cluster of collocates. A cluster is defined as the words which collocate with a given

[1] For a discussion on comparing corpora along these line on the CORPORA list, see http://www.liv.ac.uk/~tony1/corpus.html

node word (Carter, 1992: 49). Including collocates is important because they provide information about the use of key words in context. According to Sinclair 'In the relation of form and meaning it became clear that in all cases so far examined, each meaning can be associated with a distinctive formal patterning.' (Sinclair, 1993: 6). Thus, collocates can provide a better picture of the specific meanings associated to key words in context. Similarly, Phillips (1985) has found that different textbook chapters present different sets of collocations.

A problem with investigating collocations is the width of the span within which the collocates will be computed. As a general rule, a span of four words on either side of the node word is regarded as appropriate (e.g. Sinclair 1991). Sinclair, Jones and Daley, 1970 argued that 'a shorter span would miss valuable evidence, a longer one would overlay the relevant pattterns with more distant material' (:9). Nevertheless, a wider span would allow us to incorporate more context into the investigation of the contexts in which key words are used. Since the size of our data is small, we would not suffer any penalties in terms of computation efficiency if we adopted a wider span. In addition, we will only include collocates which are lexical words, so the amount of 'noise' generated by widening the span will be reduced. Therefore, we have opted for a span of five words.

3. Analysis

The business report was 3,355 words long, and the corpus, made up of 17 business reports, had 95,541 running words. The text was first tagged for part of speech by the University of Birmingham Tagger. Then the key words were computed using the WORDSMITH package (Scott, 1995), following the guidelines described earlier. A list of nouns occuring more often than expected in the text was extracted by the KEYWORDS program. After that the collocates for each key word were calculated. Only those collocates with a frequency higher than one were retained. The resulting meaning sets are listed in the table below. The item at the head of the group is the key word, and the words that follow it are its collocates. The collocates are listed in alphabetical order.

ABC, Algar and Group activities, answer, companies, company, communication, continuity, diversity, fundamentals, informatics, internationalizat

ion, investments, goals, mission, network, objective, policy, quality, performance, search, strategy, subsidiaries, treatment, trend, year

process continuity, modernization
talent human
satisfaction client, markets
objective group, priority
clients (no content word collocates)
formation coordinators, executives, program
excellence search, standards.

Three key words were joined together in the same set: ABC, Algar and Group. They were grouped because they are part of the name of the company ('ABC Algar Group'). This set is the longest of all. It comprises common themes encountered in the text: 'activities of the company', 'company investments', 'strategies deployed by the company', 'company subsidiaries', etc.

An inspection of the sets reveals that there are two types of sets. One, a central set closely associated with the company's name, where there is mention of the goals of the company, the objectives of the company, the performance of the company, the company's policies, the company's subsidiaries, the company's investments, and so on. And the other is a compound set comprising seven individual sets which express themes not formally associated with the company name. Hence, we have a *company* set on the one hand, and one *non-company* set of sets on the other. Significantly, in the non-company sets, only 4 words out of the 19, namely 'objective', 'client(s)', 'group' and 'search', are repeated in the company set. The majority are unique to the group.

We have also calculated the placement of key words across the text with the help of the KEYWORDS program. First, the program calculates the number of portions of text in which at least one key word appears. A portion is defined as 1% of the running words of the text (33.55 words). It has found that are 26 portions in which at least one key word appears. Of these, 21 contain at least one company key word. This suggests that company keywords are distributed across most of the text where key words appear. Also, it indicates that company key words share most of the portions in which noncompany key words are present. This is interesting because non-

company key words did not show as collocates of company key words or vice-versa.

By examining examples of non-company key words in the texts, we noticed that non-company key words were generally being used to refer to the company, even though the word 'company' was not mentioned. For example, the set for the key word 'process' includes the phrases 'process of developing modernization' and 'process of business modernization' which in fact should be interpreted as processes taking place *in the company*. The 'excellence' set is used in a similar context, as the example below illustrates:

> *Indispensable as tool to a company committed to modernity, Total Quality in Grupo ABC Algar has a conclusive answer. In reference to the search for excellence of operation, products, and services, through the integration of methods, efforts...*

In the example, 'search for excellence' is being used to refer to the Total Quality program taking place within the company.

These implicit references to 'the company' stand in contrast with other key words which refer to 'company' more explicitly by appearing near the word 'companies'[2]. For instance, the set for the key word 'talent' includes the phrases 'to develop Human Talents *of the companies*'.

3. Main themes

The question that remains is whether the key word sets actually represent the main themes of the text. Since the identification of themes is a highly intuitive task, a more objective means of assessment had to be devised. Summaries are usually considered representations of the main topics of the text. Obtaining an automatic summary of the text would be an adequate alternative because it would not involve subjective judgment on the part of human readers. Among the various options, automatic abridgments (Hoey, 1991) stand out as a good technique for summarizing texts because abridgments are produced on the basis of the distribution of lexical items in the text, just like our technique for extracting word sets.

[2] 'Companies' did not form part of the set because it had a frequency of one.

A series of abridgments of the business report were produced controlling for the number of links and bonds. The choice was for an abridgment which would be about 50% of the text in length, but not less. The shortest abridgment under these circumstances was one containing 1766 words and 42 sentences, representing 52.6% of the original text.

The key words and their collocates from the unabridged text were searched for in the abridgment. It was found that only one word did not appear in the abridgment ('executives'). Since the abridgment is formed by central sentences which subsume other sentences lexically, the words in the abridgment can be considered to be those which carry the main themes of the text. The fact that all but one of the set words appeared in the abridgment suggests that the sets are representative of the main themes in the text.

4. Summary and Final Comments

In this study a methodology based on the extraction of key words was applied to the identification of coherent word sets. It was hoped that these word sets would be coherent. Ultimately, it was hoped that the word sets would represent the main themes or topics of the text. Eight word sets were extracted and analyzed. It was found that the eight sets could be divided into two main groups, one 'company' set and one 'non-company' set. These two sets shared very few collocates, therefore they seem to represent distinct topics. The positions of the words in the 'company' and 'non-company' sets across the text were computed. The results indicated that they did very often shared the same portions of the text. This was interpreted as meaning that the 'non-company' sets referred to 'company' implicitly. Finally, an automatic abridgment of the text was created. The abridgment contained central sentences of the text, and therefore it was regarded as a good representation of the main themes of the text. Nearly all of the key words from the unabridged text were present in the abridgment, which suggests that the key words may indeed represent the main contents of the text.

The research reported here is work in progress and cannot offer definitive answers to the problem of identifying the main contents of texts automatically by computer. Nevertheless, it illustrates the possibility of applying key words extracted by the KeyWords program (Scott, 1995) in areas such as content analysis.

REFERÊNCIAS BIBLIOGRÁFICAS

BENSON, J. D. and GREAVES, W. S. (1992) Collocation and field of discourse. In: *Discourse description - diverse linguistic analyses of a fund-raising text.* (Eds: Mann,William C; Thompson,Sandra A) (Pragmatics and Beyond New Series, 16.) John Benjamins, Amsterdam.

CARTER, R. (1992) *Vocabulary - Applied linguistic perspectives (Reprint of 1987 original).* Routledge, London.

EGGINS, S. (1994) *An introduction to systemic functional linguistics.* Pinter, London.

HALLIDAY, M. A. K. (1992) Some lexicogrammatical features of the Zero Population Growth text. In: *Discourse description - Diverse linguistic analyses of a fund-raising text.* (Eds: Mann,William C; Thompson,Sandra A) (Pragmatics and Beyond, 16.) John Benjamins, Amsterdam, 327-358.

HOEY, M. (1991) *Patterns of Lexis in text.* (Describing the English Language.) Oup, Oxford.

KNOTT, A. and DALE, R. (1993) Using linguistic phenomena to motivate a set of rhetorical relations. Department of Artificial Intelligence, Human Communication Centre, University of Edinburgh, unpublished manuscript.

MIALL, D. S. (1992) Estimating changes in collocations of key words across a large text: A case study of Coleridge's notebooks. *Computers and the Humanities 26,* 1-12.

PHILLIPS, M. (1985) *Aspects of text structure - An investigation of the lexical organisation of text.* (North-Holland Linguistic Series, 52.) North-Holland, Amsterdam.

PHILLIPS, M. K. (1988) Text, terms and meaning: Some principles of analysis. In: *Linguistics in a Systemic perspective.* (Eds: Benson,James D; Cummings,Michael J; Greaves,William S) (Current Issues in Linguistic Theory, 39.) John Benjamins, Amsterdam/Philadelphia, 99-118.

SCOTT, M. R (1995): WordSmith Tools. Software for text analysis, unpublished computer program.

SINCLAIR, J. et al. (1970) English lexical studies. OSTI Report. University of Birmingham, Birmingham.

SINCLAIR, J. (1991) *Corpus, concordance, collocation.* (Describing English Language Series.) Oup, Oxford.

_____ (1993) Written discourse structure. In: *Techniques of description - Spoken and Written discourse* (A festsch rift for Malcolm Coulthard). (Eds: Sinclair,JMcH; Hoey,M; Fox,and G) Rout-ledge, London, 6-31.

WILSON, A. (1993) Towards an integration of content analysis and discourse analysis: the automatic linkage of key relations in text. Unit for Computer Research on the English Language Technical Papers 3, UCREL, University of Lancaster, UK.

(Recebido em maio de 1998; Aceito em novembro de 1998)