Segmenting Corpora of Texts

(Segmentação de Corpora de Textos)

Tony Berber Sardinha
(LAEL, PUC/SP)

Abstract: *The aim of the research presented here is to report on a corpus-based method for discourse analysis that is based on the notion of segmentation, or the division of texts into cohesive portions. For the purposes of this investigation, a segment is defined as a contiguous portion of written text consisting of at least two sentences. The segmentation procedure developed for the study is called LSM (link set median), which is based on the identification of lexical repetition in text. The data analysed in this investigation were three corpora of 100 texts each. Each corpus was composed of texts of one particular genre: research articles, annual business reports, and encyclopaedia entries. The total number of words in the three corpora was 1,262,710 words. The segments inserted in the texts by the LSM procedure were compared to the internal section divisions in the texts. Afterwards, the results obtained through the LSM procedure were then compared to segmentation carried out at random. The results indicated that the LSM procedure worked better than random, suggesting that lexical repetition accounts in part for the way texts are segmented into sections.*

Key-words: *Corpus linguistics; Discourse analysis; Segmentation; Lexical cohesion; Repetition.*

Resumo: *O objetivo da pesquisa apresentada é relatar um método baseado em corpus para análise de discurso que se baseia na noção de segmentação, isto é, a divisão de textos em porções coesas. Para os propósitos desse estudo, um segmento é definido como uma porção contígua de texto que consiste em pelo menos sentenças. O procedimento de segmentação desenvolvido para a pesquisa chama-se LSM ('link set median') e se baseia na identificação da repetição lexical nos textos. Os dados analisados foram três corpora de 100 textos cada. Cada corpus representava um gênero específico: artigos de pesquisa, relatórios anuais de negócio e artigos de enciclopédia. O tamanho total do corpus é 1.262.710 palavras. A segmentação por LSM foi comparada à divisão interna em seções de cada texto. A seguir, os resultados do procedimento LSM foram compa-*

rados a uma segmentação feita aleatoriamente. Os resultados indicaram que o procedimento LSM funcionou melhor do que o método aleatório, o que sugere que a repetição lexical responde em parte pela maneira pela qual os textos segmentam-se em seções.

Palavras-chave: *Lingüística de Corpus; Análise de discurso; Segmentação; Coesão lexical; Repetição.*

0. Introduction

The aim of the research presented here is to report on a corpus-based method for the analysis of discourse organization that is based on the notion of segmentation, or the division of texts into cohesive portions. By discourse organization is meant the sequential arrangement of multi-sentence units, or segments. For the purposes of this investigation, a segment is defined as a contiguous portion of written text consisting of at least two sentences (a space of text between two full stops), held together by lexical cohesive links. This follows Kukharenko (1979) and Scinto (1986), who point out that texts are constituted by sentence clusters, or 'semantic topical and lexico-grammatical unities of two or more sentences' (Kukharenko, 1979, p.235). It also ties in with a definition of text segment proposed by Fries (1995, p.54), according to whom, 'the term "text segment" is intended to apply to any chunk of text (presumably larger than one sentence in length) that is perceived as a unit'. Several authors have proposed methods for the manual analysis of discourse organization (Bronckart 1999; Georgakopoulou and Goutsos 1997; van Dijk 1997; Swales 1990, among others), but few have devised procedures that could be implemented on the computer (Bronckart 1985; Hoey 1991; Mann and Thompson 1987). The study presented here is an attempt to bridge this gap.

Corpus linguistics and discourse analysis have developed their own methods of inquiry and the contact between the two disciplines has been restricted to pragmatics and discourse markers (McEnery & Wilson 1996, pp.98-99). The analysis of discourse in the British tradition is essentially empirical (Stubbs 1996, p.23), and since corpus linguistics can also be seen as belonging to the empirical tradition of linguistic research (McEnery & Wilson 1996, p.87), there should be several ways in which the analysis of discourse could benefit from corpus-based methods. The proposal here is that one of these ways is exactly the analysis of how discourse is organized in segments.

The means whereby the segmentation is carried out on the computer is through an application of the ideas put forward by Michael Hoey in his 1991 book on the patterns of lexis in text. In his book, he analyses in detail the ways in which lexical repetition works in text and how this repetition holds the text together. The study on segmentation presented here starts from there but asks how the Hoey approach to lexis in text can be used to find the internal divisions in text. In a sense, our methodology looks at 'the other side of the coin': while Hoey was particularly interested in showing how lexis creates unity in text, we are interested in looking at lexis to show how it establishes boundaries within in the text.

1. Investigating discourse in corpora

Typically, text organization has been investigated in discourse analysis by means of the application of models which are aimed at uncovering the regularities in the constitution of the text. Models are essentially designed for hand analysis of single texts or short text fragments. The problem with discourse models, as with most linguistic theory (Mann & Thompson 1987), is that they have not been designed with computational applications in mind. Hence, discourse models are not *a priori* adaptable for computer applications.

Although computers are being used for language analysis more often, the majority of studies employing computers for language analysis are concerned with the analysis of corpora, where the interest lies not in individual texts but in collocation and word frequency. The use of computers enables the investigation of greater quantities of data; the analyses themselves are also more reliable. Using computers for the analysis of central issues in text research would allow for a better understanding of major features of texts. One particular issue which would benefit from computer-assisted analyses is text organisation.

An aspect which bears centrally on text organisation is segmentation, or the principled division of texts into constituents. For the purposes of this paper and of the segmentation procedure described here, a segment is a sequence of at least two contiguous sentences.

Segmentation is also a fundamental aspect underlying models of discourse. The research reported in this paper is aimed at developing a computer-assisted procedure for segmenting texts. In so doing, the paper aims to bring closer together corpus linguists and discourse analysts.

2. Previous approaches to segmentation

Several studies have looked at the segmentation of texts by computer. A common feature among them is the use of lexical cohesion for establishing segment boundaries (Hearst 1994, Kozima 1993, Okomura and Honda 1994). One reason for this is that, as Morris (1988, p.7) notes, 'the determination of lexical chains is a computationally feasible task'. A final feature shared by various previous approaches is the identification of lexical chains (Morris and Hirst 1991, Morris 1988, Okumura and Honda 1994).

One approach to segmentation by computer which has received considerable attention in the literature is TextTiling, a technique introduced and developed by Marti Hearst (Hearst 1994). This procedure works by identifying repetition at the word level and placing segment boundaries between paragraphs. A fundamental problem with TextTiling from the point-of-view of linguistic theory is that it operates on the principle that paragraph breaks are the only possible segmentation points in a text. This gives rise to a practical problem, in that the original segment breaks inserted by the program are then adjusted to fit the paragraph breaks. This principle restricts the number of possible places where segment boundaries can be placed, and as a result, it makes it easier for the system to place segment boundaries that match section boundaries.

Arbitrary decisions such as the restriction of segment boundary locations are not uncommon in previous approaches. For example, Hearst (1994) computed the repetition among fixed-size pseudo-sentences instead of real sentences. Youmans (1991) monitored the variation in type-token ratios in even-sized word intervals regardless of clause or sentence boundaries. And Kozima (1993) measured cohesion within intervals of a fixed length. More generally, what these studies fail to recognize is the importance of showing how messages connect across the text (Eggins 1994; Halliday 1994; Hasan 1984; Hoey 1991).

In developing the system reported on here, the first step consisted in the choice of which linguistic feature(s) to compute. Not all linguistic features that are possible to compute are relevant to text analysis, and at the same time, few relevant discourse features are computable. A review of the relevant research indicated that a feature which is both computable and closely related to how texts function (Hoey 1999) is lexical cohesion. According to Hoey, lexical repetition has a central role in the establishment of lexical cohesive links in text, and therefore lexical repetition was used as the unit of analysis. This is described in more detail in what follows.

3. Hoey's approach to lexical cohesion in text

As mentioned previously, the segmentation system developed for this research was based on Hoey's approach to the analysis of lexical relations in text. His analyses provide evidence that lexis creates cohesion mainly through repetition. As an example, consider the pair of sentences below:

(a) What is attempted in the following volume is to present to the reader a series of actual excerpts from the writings of the greatest political theorists of the past; (…)

(b) What, then, is the advantage which we may hope to derive from a study of the political writers of the past? (Hoey 1991, p.129)

These two sentences share three *links*, formed by the repetition of the following words:

Writings → writers
Political → political
Past → past

The fact that they share the three links establishes a special kind of relationship between the two sentences, namely that of a *bond*. The two sentences are bonded, that is, they can be seen as forming a pair which can be read as a single unit. It is important to point out that the two sentences, taken from a textbook introduction, were 16 sentences apart! Sentence (a) was originally sentence 1 in the text, while sentence (b) was sentence 17.

Lexis creates a range of different kinds of repetition in the text, which are the following, according to Hoey:

(a) Simple repetition: between two identical items or between items that are different because of variation in terms of simple word ending, for instance bear → bear or bear → bear*s*.

(b) Complex repetition: between similarly spelled words of different grammatical categories, or between words which share the same root, for instance human (*noun*) → human (*adjective*), or *damp* → *damp*ness.

(c) Simple paraphrase: between Two different items of the same grammatical class which are 'interchangeable in some context' (Hoey, 1991, p.69), and 'whenever a lexical item may substitute for another without loss or gain in specificity and with no discernible change in meaning' (Hoey, 1991, p.62), for instance sedated → tranquillised.

(d) Complex paraphrase: between two different items of the same or different grammatical class; this is restricted to three possibilities: (1) Antonyms which do not share a lexical morpheme (e.g. hot → cold); (2) Two items one of which 'is a complex repetition of the other, and also a simple paraphrase (or antonym) of a third' (Hoey, 1991, p.64) (e.g. a complex paraphrase is recorded for 'record' and 'discotheque' if a simple paraphrase has been recorded for 'record' and 'disc', and a complex repetition has been recorded for 'disc' and 'discotheque'; and (3) When there is the possibility of substituting an item for another (for instance, a complex paraphrase is recorded between 'record' and 'discotheque' if 'record' can be replaced with 'disc').

(e) Superordinates and hyponyms: they are considered to create repetition only if they have a common referent and if the hyponym comes first (e.g. 'bear' and 'animals' in 'a drug known to produce violent reactions in humans has been used for sedating grizzly bears . . . To avoid potentially dangerous clashes between them and humans, scientists are trying to rehabilitate the animals').

Some of the types of repetition listed above are more easily identified by computer than others; for instance, simple repetition can be spotted more effortlessly than simple paraphrase. Some kinds of repetition cannot be identified computationally at all, such as complex paraphrase. For the purposes of the investigation presented here into segmentation, simple repetition only was used. It was felt that this kind of repetition, being the most common, should enable the development of the segmentation system, without the need for other kinds of repetition. In addition, this kind of repetition can be picked up by the computer without recourse to external reference materials, such as dictionaries, thesauri and lemmatisation.

4. Adapting Hoey's approach to segment texts: The LSM procedure

The segmentation procedure developed for the present investigation is called LSM ('Link Set Median'), which is based on the identification of lexical repetition in pairs of sentences, using the Hoey scheme for lexical analysis described above.

Originally, Hoey's system was devised to show the extent of the lexical links established by repetition in texts. It also provided important insights into how sentences are bonded, forming groupings of sentence that share semantic content. This was the starting point for the segmentation procedure designed for the research reported here.

Having established that Hoey's system was useful for segmentational purposes (in addition to its purpose in showing cohesion), the next step was to make the system work to indicate the segments in the texts. To do this, a range of options was tested (see Berber Sardinha 1997). The system that worked best was the one known as the Link Set Median Procedure, or LSM.

In LSM, lexical cohesion is computed by means of counting the links that each sentence has with all the other sentences in the text. Special software was developed to locate the links in the texts. In LSM as in Hoey (1991), a link was computed whenever there was a repetition of a lexical item in a pair of sentences (Hoey 1991). Multiple occurrences of the same lexical item in any one of the sentences in the pair counted towards one link only.

Once the links between pairs of sentences were computed, link sets were identified by another specially-written program. A link set is a record of all of the sentences with which each sentence has links. For instance, if sentence 1 had three links with sentence 6 and two links with sentence 4, then its link set would be 4, 4, 6, 6, 6.

Sentence similarity was calculated by assessing the similarity between pairs of link sets. An important aspect of link sets is the fact that they allow for the identification of similarity between sentences which have no links between them. For instance, if both sentence 1 and sentence 2 had one link each with sentences 10, 11 and 12, but not with each other, then the fact that they have identical link sets could be used to reveal the extent to which they are similar.

The actual computation of the similarity between link sets was carried out through the calculation of the median for each link set. After the medians for individual link sets had been identified, the difference between each link set median and its successor in the text was tabulated. The mean of the differences for the whole text was then obtained, and it was then compared to each individual median difference. Those sentences whose median differences exceed the mean for the text are then taken to be segment boundaries. Readers are referred to Berber Sardinha (1997) for a full technical account of the procedure.

4.1. Evaluation of segmentation

The segmentation provided by the computer had to be checked against an independent criterion to see whether it was plausible. The best solution here was to check to what extent segment boundaries matched typographical section divisions in the texts. The kinds of section divisions taken into account in this investigation are those indicated by a heading (e.g. 1 Introduction, etc.).

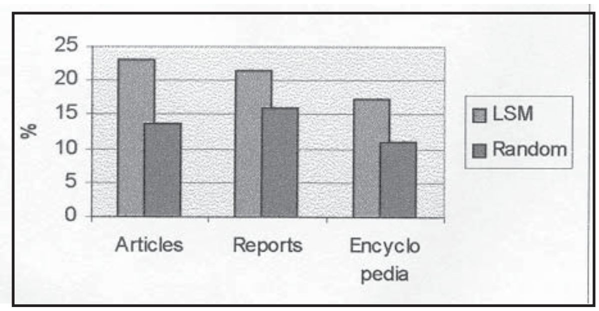
The segmentation was also checked against a random segmentation, which was obtained by having a special routine place boundaries between sentences 'blindly' across the whole corpus, that is, without any information about the lexical cohesion in the texts.

5. Corpora

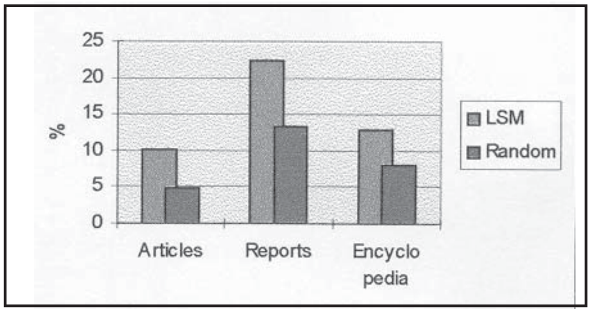
The data analysed in this investigation were three corpora of 100 texts each. Each corpus was composed of texts of one particular genre: research articles, annual business reports, and encyclopaedia entries. The size of each corpus in words was 577,026 (academic articles), 429,728 (business reports), and 255,956 (encyclopaedia entries). The total number of words in the three corpora was 1,262,710 words. The results obtained by application of the LSM procedure on the corpus were then compared to segmentation carried out at random.

6. Results

The following charts summarize the results of the application of the LSM procedure to the corpora. Two statistics were computed, namely recall and precision, based on the number of matches for each text. A match was computed whenever a boundary was inserted by LSM at a place in the text where there was a section boundary (marked by a section heading inserted by the author of the text). Recall refers to the percentage of section boundaries that matched the segment boundaries, while precision refers to the number of segment boundaries that were section boundaries.



Recall of section boundaries



Precision of segment boundaries

All differences between LSM and random rates were significant at p<.001.

The fact that LSM significantly outperformed random segmentation indicates that the segmentation approximated the section divisions in the texts.

6.1. Example

The excerpt below illustrates the actual segmentation of a text in the corpus. The excerpt runs from sentence 22 to 26, and is taken from the 'Equatorial Guinea' text (Encyclopedia corpus). There are two matches (sentences 17 and 26) and one segmentation error (sentence 22) in it. Sentence numbers appear in square brackets, and section headings are shown in italics.

=======>*Segment boundary inserted by LSM* <=======
{0017} Economy and Government

Agriculture is the main source of livelihood in Equatorial Guinea. {0018} The principal export is cacao, which is grown almost entirely on Bioko. {0019} Coffee is grown on the mainland, which also produces tropical hardwood timber. {0020} Rice, bananas, yams, and millet are the staple foods. {0021} Local manufacturing industries include the processing of oil and soap, cacao, yucca, coffee, and seafood.

=======>*Segment boundary inserted by LSM* <=======

{0022} The monetary system is based on the franc system (2864 CFA francs equal US $1; 1990). {0023} Under the 1982 constitution, the president was elected by universal suffrage to a seven-year term, and members of the legislature were elected to five-year terms. {0024} The Democratic Party of Equatorial Guinea was the sole legal political party. {0025} A new multiparty constitution was approved in 1991.

=======>*Segment boundary inserted by LSM* <=======
{0026} History

The island of Fernando Pó was sighted in 1471 by Fernão do Pó, a Portuguese navigator. {0027} Portugal ceded the island to Spain in 1778.

(Source: Encarta 1994)

A remark is in order about the segmentation error at sentence 22. The section on 'Economy and Government' is not only hybrid, as its name indicates, but parts of it could also be seen as forming colony text (Hoey, 1986, p.20). In a colony text, adjacent units typically do not form continuous prose. Sentences such as 18, 19, and 20 are loosely connected and could be read in any order. Sentence 22 marks a break in the section since there is a topic shift from 'agriculture and manufacturing' to 'currency'. The lexis in sentence 22 is very different from the preceding sentences (francs, dollars vs manufacturing, foods, etc). Thus, one could argue that sentence 22 is in fact a legitimate segmentation point and so LSM was not entirely wrong.

The full segmentation for this text is as follows:

Segment Boundary	**Section Boundary**	**Match?**
	2	No
8		No
11	11	Yes
17	17	Yes
22		No
26	26	Yes
28		No
32		No
39	39	Yes
Total: 8	Total: 5	Total: 4

In this particular text, there were 5 section boundaries in the text, and the LSM procedure placed 8 segment boundaries in it. Since there were 4 matching section boundaries, the recall rate was 80% (5/4), and the precision score 50% (8/4).

The results indicate the LSM segmentation procedure works better than random, and this suggests that lexical repetition accounts in part for how certain texts are segmented into sections.

The present levels of performance were achieved using simple repetition only. If other kinds of cohesion had been incorporated in the analysis (such as complex repetition and lexical paraphrase, or other kinds of cohesion such as substitution), it is likely that the procedure would have been more robust. In future research other types of cohesion may be incorporated in the LSM procedure to help improve its performance.

7. Conclusion

Segmentation is one kind of discourse analysis which can be carried out by computer. The discourse model which more closely corresponds to the segmentational division of discourse is the staged position (Berber Sardinha 1997). Segments resemble stages or moves (Paltridge, 1994; Swales, 1990) in that they are contiguous and sequentially arranged. The study reported here suggested that this kind of discourse analysis can be carried out across a large number of texts. It also showed that the operationalization does not need to resort to arbitrary decisions, which makes the LSM procedure valid from the point of view of linguistic theory.

The procedure developed here made use of the insights provided by Hoey's analytical system originally designed for cohesion analysis and built up from there. In the original system, sentences are joined together by lexical links and the resulting picture is one that resembles a net, with sentences sharing connections with many others in the text. For segmentation purposes, however, a net-like organization is less helpful, since the sentences in the net are all connected up in a large group. Underlying each net is a link set (not talked about as such, by Hoey), which represents all the sentences which each sentence shares links with. A net, therefore, and its associated link sets, by themselves, do not allow us to perceive the divisions in the text clearly. The link set medians, on the other hand, provided a means for estimating the similarity between pairs of link sets and for drawing boundaries within the text. Without this, Hoey's system would not have worked as a segmentation scheme. Admittedly, other measures of similarity between link sets may be employed which might bring out the similarity and differences among the link sets better than the current method, which employs medians only, and in so doing, provide a better statistic for computing segments.

The present study makes some contributions to corpus linguistics. Firstly, it helps fill a gap in the corpus linguistics literature created by a lack of discourse studies which make use of corpora (McEnery and Wilson 1996). Secondly, it shows that approaches to text linguistics and corpus linguistics are not incompatible, and therefore discourse analysis does not need necessarily to 'look within, while corpus linguistics cuts across, a text' (Renouf 1997). Finally, it indicates that it is possible in corpus linguistics to change the focus from word to text (Scott 1997), and analyse whole corpora from a discourse-centred perspective.

Recebido em agosto de 2000.
Devolvido pelo autor em outubro de 2002.

References

Berber Sardinha, A. P. 1997. Automatic identification of segments in written texts. Unpublished PhD thesis. Liverpool: University of Liverpool. Available online at lael.pucsp.br/~tony.

Bronckart, J. P. 1985. *Le Fonctionnement des Discours – Un Modèle Psychologique et un Méthode D'Analyse*. Neuchâtel, Paris: Delachaux & Niestlé.

Bronckart, J. P. 1999. *Atividades de linguagem, discursos e textos*. São Paulo: EDUC.

Fries, P. H. 1995. Patterns of information in initial position in English. In *Discourse in Society: Systemic Functional Perspectives* (Meaning and Choice in Language: Studies for Michael Halliday) Peter H. Fries and Michael Gregory (eds.). Norwood, NJ: Ablex: 47-65.

Georgakopoulou, A., & Goutsos, D. 1997. *Discourse Analysis – An Introduction*. Edinburgh: Edinburgh University Press.

Halliday, M. A. K. 1994. *An Introduction to Functional Grammar*. 2nd ed. London: Edward Arnold.

Hasan, R. 1984. Coherence and cohesive harmony. In *Understanding Reading Comprehension: Cognition, Language and the Structure of Prose*. J. Flood (ed.) Newark, Delaware: International Reading Association: 181-219.

Hearst, M. A. 1994. Multi-paragraph segmentation of expository texts. Project Sequoia Technical Report 94/790. University of California at Berkeley.

HOEY, M. 1991. *Patterns of Lexis in Text*. Oxford: Oxford University Press.

KOZIMA, H. 1993. Text segmentation based on similarity between words. Unpublished manuscript, University of Electro-Communications, Tokyo, Japan.

KUKHARENKO, V. 1979. Some considerations about the properties of texts. In *Text vs Sentence*, J. PETOFI (ed.) Vol. 1. Hamburg: Helmut Buske Verlag: 235-45.

MANN, W. C., & THOMPSON, S. 1987. Rhetorical Structure Theory: A theory of text organization. Manuscript. ISI Reprint Series 87/190.

MCENERY, T., & A. WILSON 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

MORRIS, J., & G. HIRST. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics* **17**: 21-48.

MORRIS, J. 1988. Lexical cohesion, the thesaurus, and the structure of text. Technical Report 219. University of Toronto, Toronto.

OKUMURA, M., and T. HONDA. 1994. Word sense disambiguation and text segmentation based on lexical cohesion. Paper presented at COLING 1994.

PALTRIDGE, B. 1994. Genre analysis and the identification of textual boundaries. *Applied Linguistics,* **15**: 288-299.

RENOUF, A. 1997. Teaching corpus linguistics to teachers of English. In WICHMANN, A., FLIGELSTONE, S., MCENERY, T., and KNOWLES, G. (eds) *Teaching and language corpora*: London: Longman: 255-266

SCINTO, L. F. M. 1986. *Written Language and Psychological Development.* Orlando, Fla: Academic Press.

SCOTT, M. 1997. PC Analysis of key words – and key key words. *System* **25**: 233-45.

STUBBS, M. 1996. *Text and Corpus Analysis – Computer-Assisted Studies of Language and Culture*. Oxford: Blackwells.

SWALES, J. M. 1990. *Genre Analysis – English in Academic and Research Settings.* Cambridge: Cambridge University Press.

VAN DIJK, T. 1997. *Discourse Studies: A Multidisciplinary Introduction*. (Vol. 1 – Discourse as Strucuture and Process London: Sage.

YOUMANS, G. 1991. A new tool for discourse analysis: The Vocabulary Management Profile. *Language* **67**: 763-89.