D.E.L.T.A., Vol. 12, N° 1, 1996 (35-56)

LEXICAL COHESION AND THE REPETITION MATRIX*

Antonio Paulo BERBER SARDINHA (Doutorando, English Department, University of Liverpool, England)[1]

ABSTRACT: O estudo da coesão lexical tem sido comumente feito em relação à questão da textura. Entretanto, a coesão lexical também é capaz de revelar outra propriedade textual: a segmentabilidade, isto é, como um texto é formado de várias partes interconectadas simultaneamente. Uma técnica de investigação da coesão lexical que pode ser reutilizada para a investigação da segmentabilidade é a 'análise de adesão' ('bonding analysis'). Neste tipo de análise são descritos os vários laços coesivos que se formam através de todo o texto, com a possibilidade da manipulação do grau de adesão. Ademais, os resultados da análise de adesão são mostrados em uma matriz numérica, o que torna possível a visualização da distribuição dos laços coesivos do texto. Neste trabalho, estas propriedades da análise de adesão são canalizadas para a análise de segmentabilidade através de um procedimento chamado de 'faixa de proximidade' ('vicinity strip'). O procedimento é descrito, aplicado, e os resultados são discutidos.

Palavras-chave: léxico; coesão; adesão; texto; segmentação.

Key-words: lexis; cohesion; bonding; text; segmentation.

0. Introduction

The study of lexical cohesion has usually been seen in relation to texture (Halliday 1985). As such, the patterns formed by lexical cohesive ties across texts partly explain why portions of language are texts and not simply a random selection of words. However, lexical cohesion also reveals other important properties of language besides texture. One of these properties is segmentability (Hearst 1994; Kozima, 1993), which means how texts are seen not as a continuous whole but as a complex grouping of larger pieces.

One analytical technique which can be used to investigate segmentability by means of lexical cohesion is bonding analysis (Hoey, 1991). In bonding analysis, a comprehensive mapping of concurrent widespread lexical cohesive ties is implemented, which results in a thorough description of how texts hang together, as well as opening up the possibility for explaining how texts segment.

The main aim of this study is to show how to use bonding analysis for text segmentation. In order to accomplish this aim, an original technique will be introduced: the *vicinity strip*. Because the vicinity strip is a consequence of a particular method of interpretation of the repetition matrix found in bonding analysis, it is necessary to introduce how bonding analysis is carried out. A major section of the paper will thus be devoted to discussing bonding analysis, even though introductions to bonding analysis per se can be found elsewhere (Berber Sardinha, 1993a,b).

1. Lexical cohesion and segmentation

The relevance of the use of bonding analysis in text segmentation lies in the need to make clear how topics in texts are created and distributed by means of lexical cohesive ties. The demarcation of topics in texts has been seen as the sequential computation of local features of lexical distribution and concentration (Hearst, 1994; Kozima, 1993). Unlike these approaches, in this paper emphasis is put on how simultaneous threads of lexical ties are formed across the whole text. Nevertheless, such a picture of how lexical cohesion operates generates a much more complex rendering of the lexical formation of topics. Therefore, a different approach to text segmentation is needed to enable the analyst to demarcate topic boundaries while recognizing the breadth and reach of lexical connections within the text. Such an approach is offered here.

There have been indications that lexical cohesion reveals topic segmentation (Phillips 1989, Hasan 1984). Phillips (1989) refers to topics by means of 'lexical clusters'. In his study, he seeks for clusters of words in science textbooks by identifying patterns of co-occurrence. He uses statistical analysis to identify collocational groups that recur across different chapters of the textbooks.

Intercollocation patterns are used to generate 'lexical macrostructures', which correspond to the contents of the chapters of the textbooks.

Hasan's (1984) approach is based on the notion of lexical chains in text. The interaction of chains indicates in some way the existence of different blocks of topical information in the texts.

There are other studies outside discourse analysis which have concentrated on text segmentation (Hearst, 1993, 1994; Hearst and Plaunt, 1993; Kozima, 1993; Kozima and Furugori, 1993; Morris, 1988). These studies indicate that lexical cohesion is mapped onto text segments, and therefore, the identification of lexical cohesive patterns leads to demarcation of text topics. Because these studies do not focus on how language works, but rather on how a computer application models some specific feature of language and how the application output is useful to solve practical problems, they are not seen as primary references for the present paper.

2. Bonding analysis

Hoey's (1991) bonding analysis is a scheme for identyfing semantically-related portions of text. In practice, in bonding analysis one identifies pairs of sentences which either say the same thing or refer to the same thing. His approach is not specifically geared to the identification of topics. Rather, it is a scheme for identifying the multitude of lexical connections that are built in natural texts. However, bonding analysis does not indicate an even distribution of lexical links across the text; it shows, rather, that there are concentrations of lexical content in some parts of the text. Since one of the consequences of bonding is that portions of text make sense together, it is possible to look for topical segments by identifying concentrations of bonded sentences.

3. Segmentation by clustering

The basis for the identification of segments is the occurrence of 'clusters'. By cluster it is meant a group of sentences that are linked by a certain number of repetitions and that make sense when read

together. The following is an example of a cluster (see excerpt 1).

Excerpt 1 Example cluster

41 In the area of TELEPHONE TERMINALS, we < ACME > concluded two very important projects of products lacking in our < ACME > catalogues: the public coin-payphone and the mini-telephone.
42 The first one < public coin-payphone > represents an innovation in design, being modern and resistant.
43 The payphones in a district can be supervised by a centralized system which allows the operating company to control their < public coin-payphones > integrity, income, alarms and addresses by means of a central computer and a specific software.
44 We < ACME > believe that the features of this product < public coin-payphone > meet the expectations of all operating companies who aim at a safer and more efficient operation.
45 The mini-telephone represents a great achievement in terms of miniaturization, as the whole telephone has been assembled in a single piece of attractive design and colours.
46 Due to its < mini-telephone > versatility, this product < the mini-telephone > can be used either as a desk or a wall telephone and the user, by means of a small key, can switch it for decadic or touch signalling.
47 Both products < the mini-telephone and the coin-operated payphone >, manufactured by ACME AMAZONIA will be liable in the first quarter of 1992.

For the kind of analysis pursued here, it is vital to identify as belonging to the same cluster sentences that occur in succession at a certain distance from each other. It is believed that the more adjacent sentences are, the more likely it is that they will make sense together, and by extension, the more likely it is that they will form a cluster.

3.1. Links and bonds

As said above, a link is a relation between two words in two

different sentences. A bond, in turn, is a relation between two sentences that have a certain number of links. The smallest number of links that configure a bond is set at three (Hoey 1991).

For instance, sentences 1 and 8 have no link (see excerpt 2):[2]

Excerpt 2 Sentences 1 and 8 with zero link

1 1991 clearly indicated the trend, already evident in past years, towards the need of a new model for the industrial companies, enabling them to work efficiently under extreme competition and in a market where the driving forces, free from governmental intervention and protectionisms, merged with new technologies and industrial processes, demanded changes in strategies and attitudes.
8 The tasks not related to the core business were moved to outside suppliers.

In order to determine whether two sentences are bonded or not, we must decide on the lowest number of links at which bonding can be said to occur. The smallest number of links is arbitrarily set to 3. Three links is referred to as 'cut-off' 1; four links as 'cut-off' 2, and so on. The higher the cut-off point, the more we constrain the chances for bonding.

If we operate at the minimum cut-off point 1, or 3 links, for instance, we can consider sentences 1 and 15 as bonded (see excerpt 3).

Excerpt 3 Sentences 1 and 15 with 3 links

1 1991 clearly indicated the trend, already evident in past years, towards the need of a new model for the industrial companies, enabling them to work efficiently under extreme competition and in a market where the driving forces, free from governmental intervention and protectionisms, merged with new technologies and industrial processes, demanded changes in strategies and attitudes.
15 Important changes in managers' attitudes have also been

achieved, amongst which the pro-active and entrepreneurial actions should be stressed, as well as taking over risks and responsibilities, following < ACME > values, ie, professionalism (action in a customer's oriented way and towards the company < ACME > goals), respect (team-work with mutual confidence and effective cooperation besides showing respect to others) and perseverance (continue to pursue ACME goals).

Links 1 attitudes → attitudes
2 changes → changes
3 companies → ACME,company,ACME,ACME

If the cut-off point for bonding is raised from 1 (3 links) to 3 (5 links), the previous pair of sentences 1 and 15 would not be bonded anymore; sentences 1 and 4, on the other hand, would (see excerpt 4).

Excerpt 4 Sentences 1 and 4 with 5 links

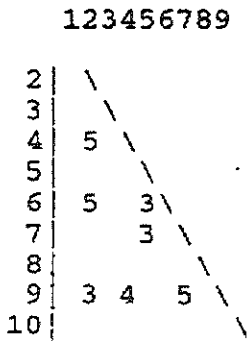
1 1991 clearly indicated the trend, already evident in past years, towards the need of a new model for the industrial companies, enabling them < the industrial companies > to work efficiently under extreme competition and in a market where the driving forces, free from governmental intervention and protectionisms, merged with new technologies and industrial processes, demanded changes in strategies and attitudes.
4 The new technologies and their < new technologies > applications also demand higher efficiency and greater resources in the after sales services, which constitute a differentiation element among competitors.

Links 1 competition → competitors
2 demanded → demand
3 efficiently → efficiency
4 new,new → new,new
5 technologies → technologies,technologies

3.2. Repetition matrix

In bonding analysis, the links are recorded on a symmetrical matrix. This matrix is made up of two axes indicating the positions for the sentences in the text, and is sliced up in half from two opposing corners to avoid redundant positions. Let us look at table 1.

TABLE 1 Example matrix



In table 1, the numbers from top to bottom down the left margin and from left to right at the top stand for sentences 1 to 10. The digits within the space between the diagonal and the vertical lines are the number of links shared by a particular sentence pair. So, from top to bottom, the first number '5', for example, indicates 5 links between sentences '4' and '2'; the first number '3' represents 3 links between sentences '6' and '5', and so on.

3.3. Vicinity strip

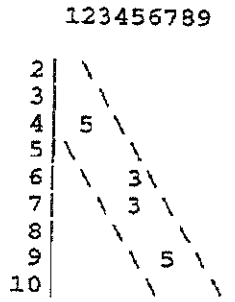
The vicinity strip is a technique that originates from a particular distributional feature of lexical cohesion as revealed by bonding analysis, namely that the number of links shared by sentence pairs tends to increase as the distance between sentences decreases. In representational terms, this means that there is a greater concentration of figures indicating links near the diagonal of the matrix. Since concentrations of links are potential indicators of clusters, establishing a maximum distance away from the diagonal

within which links can form clusters, can lead to segmentation. The 'vicinity strip' does exactly this: to exclude an area from the matrix for analysis in clusters.

In short, the vicinity strip is a means of restricting the choices of bonded sentences that can form clusters by reducing the distance at which sentences can bond.

Therefore, in looking for clusters with the 'Vicinity Strip' technique, one draws a parallel line at a certain distance from the leading diagonal in the matrix. Using the example matrix that we showed before, we can draw an arbitrary 'vicinity strip' to represent only those pairs of sentences that are at most 3 sentences apart (see table 2).

TABLE 2 Example vicinity strip



4. Methodology

Before attempting to draw the vicinity strip for a whole text, one must follow several steps. The whole technique, including the drawing of the strip, is not based on straightforward mathematical calculations, but rather on decisions based on inspection of the various matrices formed at various cut-off points.

The steps in the analysis according to the 'Vicinity Strip' technique can be sketched as follows (see table 3).

TABLE 3 Steps

1	Examine repetition matrices at each cut-off point
2	Decide on a cut-off point for inclusion in clusters
3	Decide on a vicinity range
4	Draw strip
5	Draw continuity graph
6	Mark off clusters
7	Interpret clusters
8	Evaluate vicinity strip clustering ability

5. Text analyzed

The text analyzed in this study is an 83-sentence business report written in English for a multi-national telephone company that operates in Brazil (due to a lack of space, the text is not reproduced here; see Berber Sardinha 1993a). Because the analysis was done by computer, several modifications had to be made in the text to allow the computer program to identify the links correctly, including the restoration of the references of all referents (for more details, see (Berber Sardinha, 1993a).

6. Results and interpretation

In what follows we will present the results of the application of the technique presented above.

7. Examination of cut-off points

The reader will be referred to look at the matrices in the appendix. Due to a lack of space, not all matrices will be displayed (see Berber Sardinha 1993a for the complete set of matrices). Starting with matrix 3 for cut-off level 5 (8 links) in appendix 3, up to matrix 1 for cut-off level 1 (3 links) in appendix 1, there are two trends which will affect the position of the vicinity strip. First, there is a decrease in the number of bonded sentences as the cut-off points become higher. Second, there are more bonds between sentences far

from each other as the cut-off point is reduced. This can be seen by comparing matrix 2, for cut-off point 4, to matrix 1, for the lowest cut-off point possible. These observations indicate that a higher cut-off point not only excludes many lexically-cohesive sentences, but it mainly excludes cohesive sentences that are away from each other. The vicinity strip fits perfectly into this description of the distribution of links across different cut-off points.

In addition to distributional evidence, one has to look for textual evidence that clusters are well-formed near the diagonal, thus being capturable by the vicinity strip. A possible cluster can be seen in matrix 1 beginning at sentences 33 and 34 (sse example 5). In table 4 below, we depict this sequence of bonded sentences in italics. The bonded sentences in the maximum range in this cluster, which may serve as our vicinity strip, is shown in italics.

Excerpt 5 Sentences in provisional vicinity range

33 New products are being brought to the market to comply with the demand for mobility, anticipating the so-called "personal commu-nications network", where the CELLULAR MOBILE TELEPHONY holds a strategic position.
34 It < CELLULAR MOBILE TELEPHONY > offers the advanced facilities of the fixed telephone network, added to the extraordinary advantage of user mobility.
35 It < CELLULAR MOBILE TELEPHONY > consists of one or more Switching Mobile Centers (SMC) to which the Radio Base Stations (RBS) are connected to support the radio channels covering the service areas.
36 A variant of mobile communications to be used by private networks is the TRUNKING.
37 It < Trunking > is a radio system which enables a mobile user to communicated briefly (for dispatches with an average duration of 20 seconds) with other users within the same system or, through the public telephone network, with ordinary subscribers.
38 The good performance of the CELLULAR MOBILE TELEPHONY, MOBITEX, TRUNKING and WIDE AREA

positioning and distribution of the Radio Base Stations,which will cover the areas where the user is moving in.

39 Aiming at calculating accurately the position of these < Radio Base Stations > units before installing the system, ACME has equipped its < ACME > cellular planning center, in any city, with a powerful software which analyses the radio propagation as a function of the area topography.

40 This CAE (Computer Aided Engineering) < software > was licensed from ACME Radio Systems AS and will be continuously updated in order to cope with the new facilities added to the system < of CELLULAR MOBILE TELEPHONY, MOBITEX, TRUNKING and WIDE AREA PAGING >.

TABLE 4 Provisional width of vicinity strip

```
          1111111111222222222233333333333
12345678901234567890123456789012345678901
34                          33      4\
35                                   3\
36                                  33 \
37                                  44 5\
38                        3         34934\
39                    4               6  6\
40                  33 33             455 486\
```

This excerpt reads very coherently, and therefore qualifies as an optimal distance at which to place the vicinity strip. The maximum width in this cluster is 7 sentences, therefore we will use seven as our vicinity range.

8. Drawing of strip

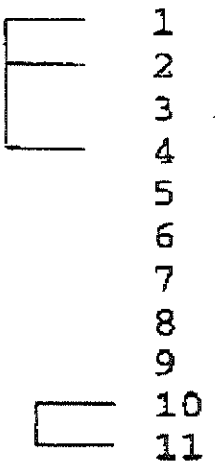
Having decided on a vicinity range, we can now draw a vicinity line that parallels the leading diagonal in the matrix at the specified distance. Besides, for the sake of ease of reference, we should erase those bonded sentences that fall out of the strip. That can be seen in appendix 4.

9. Drawing of continuity graph

Now that we have the sentence pairs that occur inside the strip, we must represent the bonds between each pair graphically before we can see where each cluster ends and another one begins. We will do this by simply arranging the sentences in sequence and drawing a line from one sentence to another, in what we may tentatively call a 'continuity graph'.

We begin by drawing a line from sentence 1 to sentence 2, as these are the first ones to appear in the matrix where the strip is demarcated (appendix 4). Then we draw a line from sentence 1 to 4, then from sentences 10 to 11. At this point we have the following (see table 5).

TABLE 5 Continuity graph for initial sentences



A consequence of this two-dimensional representation of bonded sentences is that pairs which are not bonded will appear as if they were. Interestingly enough, such pairs tend to show connections to the cluster by processes such as 'bridging' and 'bracketing' (cf Berber Sardinha, 1993a).

10. Marking off clusters

We draw the boundaries between clusters by graphically connecting the first and last sentences that are joined by a line in the continuity graph (see appendix 5). Note that we leave out several sentences in between the first and last continuously joined sentences.

The clusters that we have found out comprise the space between the sentences (see table 6).

TABLE 6 Sentences in clusters

	Sentences
1	1 to 4
2	7 to 16
3	18 to 20
4	21 to 24
5	25 to 32
6	33 to 40
7	41 to 47
8	49 to 53
9	60 to 61
10	63 to 69
11	71 to 73
12	74 to 75
13	77 to 78
14	82 to 83

11. Interpretation of clusters

We begin by looking at how the clusters that identified are mapped onto the existing boundaries across text sections. This is represented in the continuity graph in appendix 5.

Overall, we notice that there is a mismatch between cluster and section boundaries. The exceptions are clusters 3 and 14, which correspond to the exact area of the sections 'Previ-ACME' and 'Report of Fiscal Year', respectively. We also notice that some clusters split a section in nearly two halves: the 'New Products' section is cut up between clusters 6 and 7. Conversely, there are clusters that reach across more than one section; it is the case of cluster 2, which takes portions of both 'Plan of changes' and 'Human Resources', and cluster 5, with links 'Domestic market', 'International market' and the beginning of 'New Products'. There are also sections that are not represented by any cluster, such as 'Acknowledgments', which is made up of three isolated sentences. In general, there are more clusters than there are sections, and the clusters do not map all of the sentences in the text. We notice several gaps in the continuity graph, especially at the bottom, where we see a sequence of 2-sentence clusters.

Now we will look at how each cluster is formed. The first one, from sentences 1 to 4, emphasizes the need for change in terms of more efficiency and the relation between new technologies and the need for efficiency. This cluster cuts across the 'Letter to shareholders' sections in which sentence 1 is located and the 'Plan of changes' section. Cluster 2 begins at sentence 7 and stretches up to sentence 16, and focuses on the notion of need for changes in the human component of the company. This is why it spans across the sections 'Plan of changes' and 'Human resources.' This cluster is characterized by links of lemmata like 'emphasis', 'change', 'responsibility', 'management', 'job' and 'rotation', which signal the intention of changing attitudes among the managerial personnel of the firm. Cluster 3 comprises sentences 18 to 20, and matches the section 'Previ-ACME', therefore needing no validation. Cluster 4 is initiated by sentence 21 and stretches up to sentence 24. In this cluster the prevailing lexis centers around the notion of digital

exchange equipment and the company's participation in this market segment.

Cluster 5, from sentences 25 to 32, spans across three sections: about half of 'Domestic market', the whole of 'International Market' and the beginning of 'New products'. As we examine this cluster, we notice that the section 'International Market' is locked up inside the cluster between bonded sentences 25 and 32. Sentences 28 and 29 of 'International Market' do not present any links with the sentences other than 28 and 29. It therefore characterizes what we may call an 'inset' in the cluster. What happens here is the opposite to the possibilities of inclusion in the cluster that we termed 'bridging' and 'bracketing'. An 'inset', on the other hand, is an exclusion from the cluster; a separate portion that forms a cluster by itself, but which does not break up the cluster that surrounds it.

Cluster 6 stretches from sentences 33 to 40 and occurs inside the 'New products' section. This cluster identifies the portion of the 'New products' section that revolves around the notion of 'cellular mobile communications'. Cluster 7, from sentences 41 to 47, immediately follows cluster 6 and appears at the end of the 'New products' section, focusing on the 'mini-telephone' and the 'coin-operated payphone'. The technique correctly identified the heterogeneity in the 'New products' section by identifying two contiguous clusters in what was signalled in the text as a larger whole comprising two separate blocks. Cluster 8 occurs a few sentences away from the end of cluster 7, and takes the initial third of the 'Technology' section; this cluster mainly deals with the 'AXE digital exchange'.

Between clusters 8 and 9 there is a zone from sentences 54 to 59, comprised by 3-link bonds, that was not captured in the vicinity strip at cut-off point 2, which we settled for. This low-linkage cluster focuses on the 'technical training center'. Cluster 9, which takes up only a couple of sentences (60 and 61) inside 'Technology', is formed by lexis referring to 'complementary methodology of training' and the 'computer training center'. Cluster 10, from sentences 66 to 69, stretches across 'Destination of net profit' and the first sentence of 'Increase in capital'. It centers around the reference to 'increase' at the 'Annual General Meeting' of the company; in the cluster, mention is

made to an increase in quality and in capital. Cluster 11 appears at the final half of 'Increases in capital', from sentences 71 to 73, and is made up of the repetition of the 'nominal value of each share'.

The remaining clusters are pairs of sentences. Cluster 12, sentences 74 and 75, occurs at the beginning of 'Consolidated and affiliated companies', whereas cluster 13, sentences 77 and 78, closes off this section. The former cluster focuses on the reduction of operations of ACME Amazônia, while the latter refers to the transfer of capital from one of ACME's competitors. And finally, cluster 14, with sentences 82 and 83, which correspond to the whole of the 'Report of Fiscal Year' section.

12. Final remarks

One of the positive aspects of the vicinity strip is that it is based on the observation of how lexical cohesion is distributed across the text without predefined values. The vicinity width is set according to observation of how links are affected by different bonding cut-off points. This adds to the validity of the procedure.

One of the weaknesses of the procedure is that some portions of the text are left out of the resulting segments. This is due to the fact that some portions are not lexically cohesive, and not actually because the procedure is faulty. Future research is needed to investigate the influence of different cut-off points in clustering.

(Recebido em 10/06/1994. Aceito em 18/11/1994.)

NOTES

* An expanded version of this paper appears in Berber Sardinha (1993a).

1 Bolsista CNPq - Brasília, Brasil

2 The words in '<>' refer to words added during the restoration of the references in the text. For an explanation of restoration see 'Text handling' below.

REFERENCES

BERBER SARDINHA, A. P. (1993a) Lexis in annual reports: the vicinity strip technique. Working Paper 1. *DIRECT Papers*. CEPRIL, PUC-SP, São Paulo, SP, Brazil/AELSU, English Dept, Liverpool University, UK,.[Available by anonymous ftp, from ftp.liv.ac.uk, directory ~ftp/pub/linguistics, file wp-1.ps.gz, or http://www.liv.ac.uk/~tony1/linguistics.html].

BERBER SARDINHA, A. P. (1993b) Lexis in annual reports: the cluster triangle technique. Working Paper 2. *DIRECT Papers*. CEPRIL, PUC-SP, São Paulo, SP, Brazil/AELSU, English Dept, Liverpool University, UK,.[Available by anonymous ftp from ftp.liv.ac.uk, directory ~ftp/pub/linguistics, file wp-2.ps.gz, or http://www.liv.ac.uk/~tony1/linguistics.html].

HALLIDAY, M. A. K. (1985) *An Introduction to Functional Grammar*. London: Edward Arnold.

HASAN, R. (1984) Coherence and cohesive harmony. In J. FLOOD (ed) *Understanding Reading Comprehension*. Delaware: International Reading Association.

HEARST, M. A. (1993) Texttiling: a quantitative approach to discourse segmentation. *Project Sequoia technical report 24/93*. [Available via ftp from cs-tr.cs.berkeley.edu].

HEARST, M. A. (1994) Multi-paragraph segmentation of expository texts. *Project Sequoia technical report 94/790* [Available via ftp from cs-tr.cs.berkeley.edu].

HEARST, M. A., and C. PLAUNT (1993) Subtopic structuring for full-length document access. *Project Sequoia technical report 93/26* [Available via ftp from cs-tr.cs.berkeley.edu].

HOEY, M. (1991) *Patterns of Lexis in Text*. Oxford: OUP.

KOZIMA, H. (1993) Text segmentation based on similarity between words. *Proceedings of ACL-93*: 296-8.

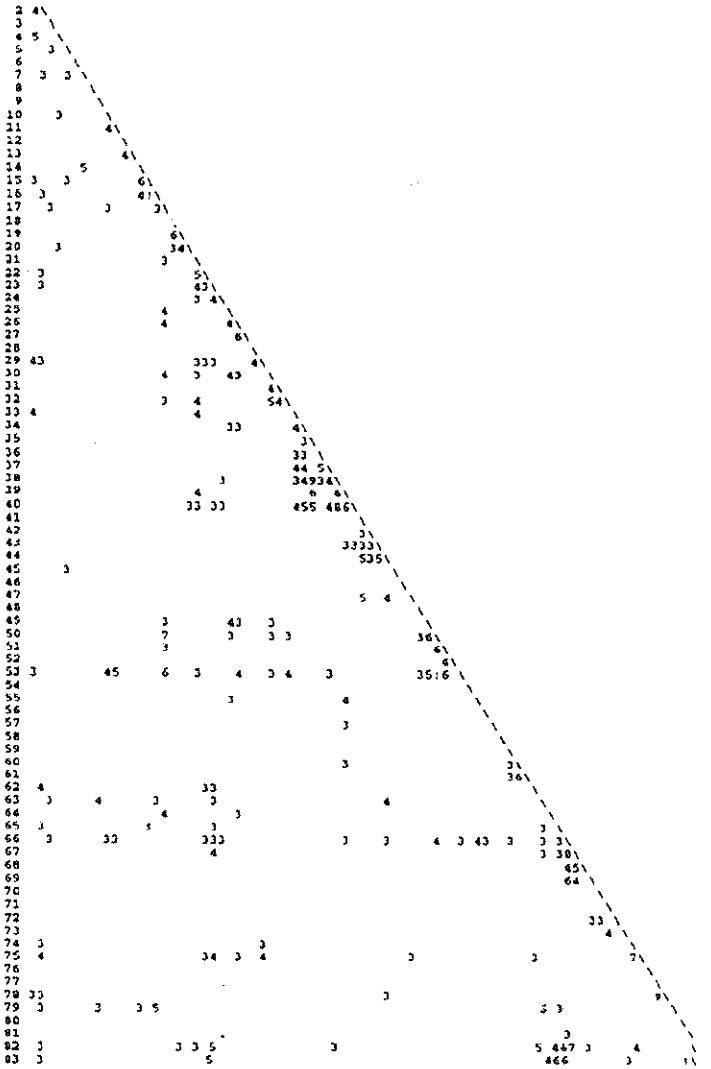
KOZIMA, H. and T. FURUGORI (1993) Segmenting narrative text into coherent scenes. *Proceedings of EACL-93:* 232-9.

MORRIS, J. (1988) Lexical cohesion, the thesaurus, and the structure of text. *Technical Report CSRI-219*. University of Toronto.

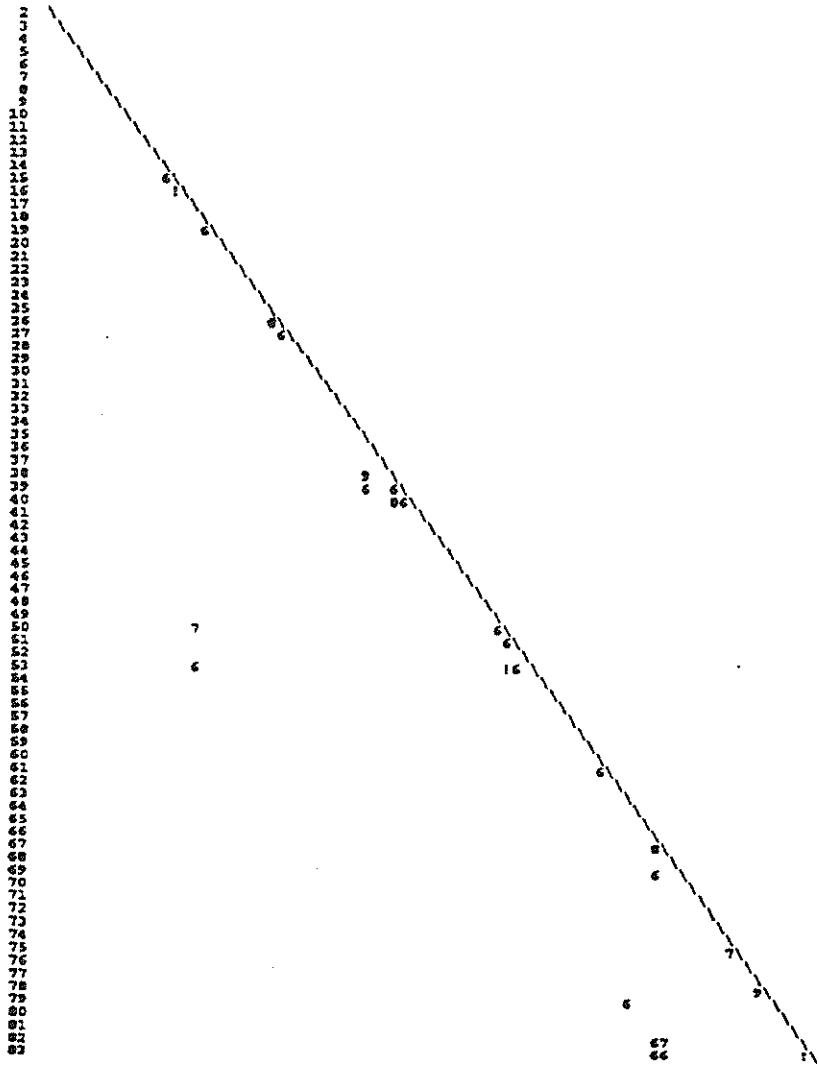
PHILLIPS, M. (1989) *Lexical Structure of Text*. (Discourse Analysis Monographs 12). Birmingham: ELR, University of Birmingham.

APPENDICES

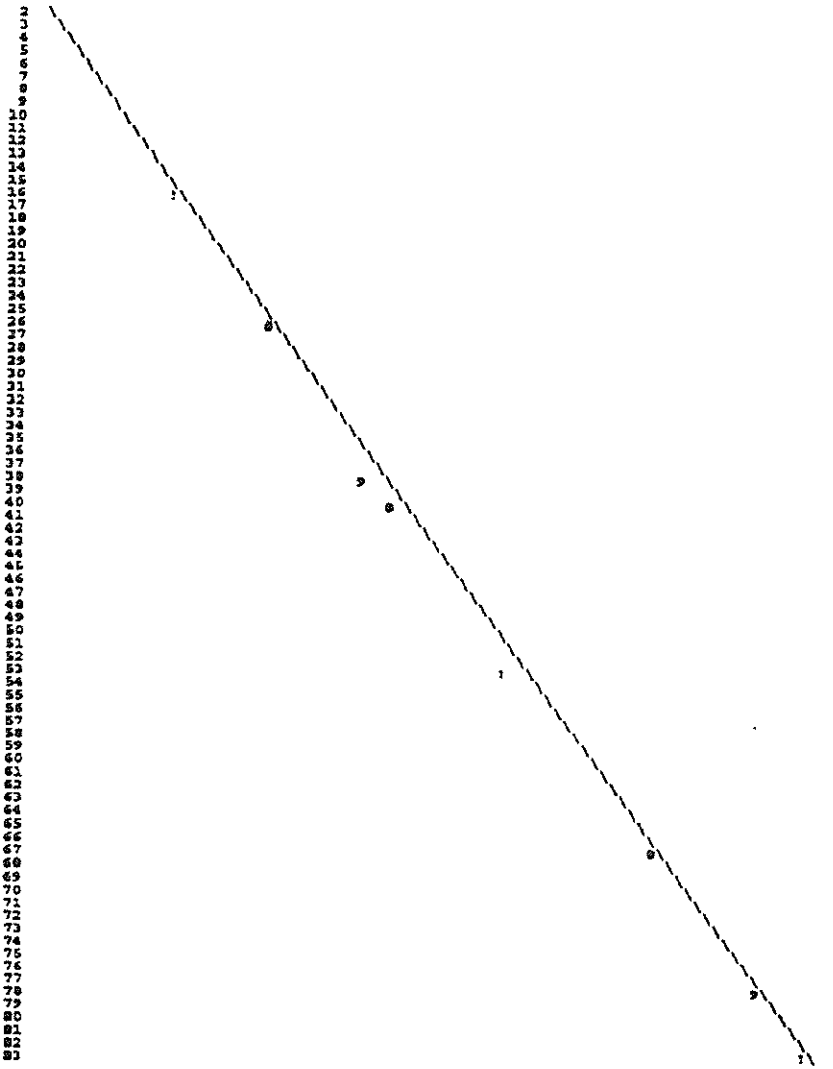
1 Matrix for cut-off point 1, 3 to 12 links



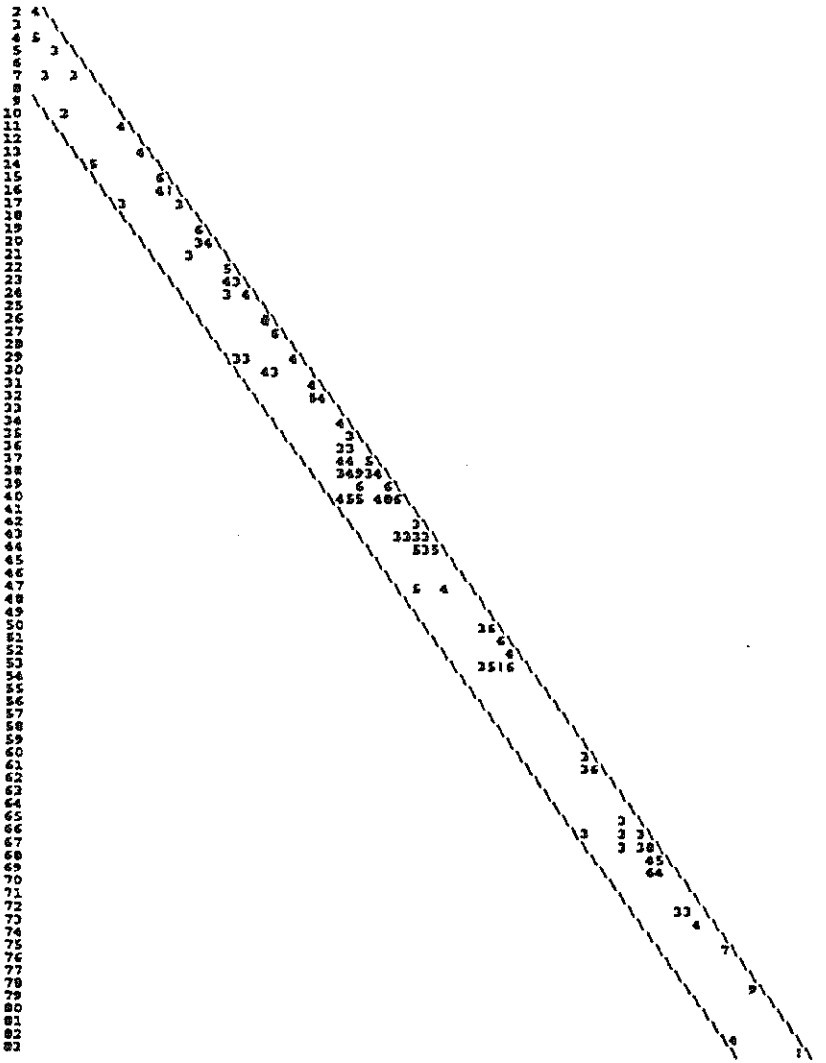
2 Matrix for cut-off point 4, 6 to 12 links



3 Matrix for cut-off point 6, 8 to 12 links



4 Vicinity strip width 7, at cut-off point 2=4 to 12 links



5 Continuity graph with clusters and text sections