On the Generality of Thesaurally derived Lexical Links

Jeremy Ellman and John Tait

School of Computing, Engineering and Technology, University of Sunderland,
St Peter's Basin
Sunderland SR6 0DD
UK
Jeremy.Ellman@sunderland.ac.uk

Abstract

Cohesion is that property of a text that allows it to be read as a unified entity rather than a series of unconnected sentences. Lexical cohesion may be detected using an external thesaurus and the resulting representation used in a variety of language processing tasks. Our particular interest is in determining whether texts of different genres are similar in meaning. For this, we wish to derive a measure based on lexical cohesion. Consequently, we need to determine if lexical cohesion is independent of genre or a function of it.
This paper examines the statistics of lexical cohesive relations. Our method involves determining the distribution of lexically cohesive relations in several book length texts. These are shown to have different reading complexities, but equivalent cohesive properties. From this, we conclude that lexical cohesion is independent of reading complexity.

Keywords: Lexical cohesion statistics, lexical chains, readability metrics, genre.

1. Introduction

Cohesion is that property of a text that allows it to be read as a unified entity, as opposed to a series of unconnected sentences. For example, a scientific paper will have logical argument followed by evidence to support it. Halliday and Hasan (1976, 1989) have identified many devices used to make text cohesive. These include linguistic phenomena, such as anaphora, cataphora, ellipsis, co-extension, and words linked into chains. These 'lexical chains' may be composed of identical or similar words.

Morris and Hirst (1991) proposed using Roget's thesaurus to identify the lexical chains in a text. This would suggest the texts' cohesive structure, which is an essential step in determining its deeper meaning. Lexical chains may be identified using an external thesaurus such as Roget's, or WordNet (Miller 1991). Lexical chains have been applied to several different areas of language processing such as word sense disambiguation and text segmentation, (Okumura and Honda 1994), malapropism detection (StOnge 1995), detection of HyperText links in newspaper articles (Green 1997), and Information Retrieval (Stairmand 1996, Smeaton 1999).

The lexical chaining approach to text analysis is highly attractive, since it is both robust, and deals with whole texts. It is though a heuristic approach, and there are a number of unanswered questions associated with it. In particular, we do not know whether it is a reflection of a text's genre, or an independent measure.

Our particular interest is in the use of lexical cohesion to define a measure that may be used to determine whether two or more texts are similar in meaning. This has applications in both information retrieval and textual case based reasoning (Smeaton 1999, Lenz 1998).

First, we have to question the techniques' general applicability. It may be that links between words, or lexical links, have different properties depending on the type of text. The lexical chains derived from the links would then be some measure of reading complexity, rather than a general technique.

Next, we need to consider the text window within which words should be considered for linkage. Clearly the greater the number of words considered, the more time consuming the algorithm will be. Morris & Hirst (1991) check for identical words throughout the whole text, whereas they only look for lesser relationships within a four-sentence distance. This is similar to the fifty-word window Yarowsky (1992) used in his work on lexical disambiguation.

Finally, the effectiveness of Morris and Hirst's different word linking relationships have not been tested. Thus, it maybe that some complex relationships are hardly ever found in real texts, and may be ignored in practice.

To answer these questions, we decided to analyse a set of longer texts. Since most lexical chaining has considered shorter texts, results, though interesting, may not be general. Consequently, a mixed range of texts of differing complexity were chosen. These varied from children's books such as "Alice in Wonderland" to more challenging works such as Kant's "Critique of Pure Reason".

This paper proceeds as follows: Firstly, we discuss the selection of a set of texts for our initial investigation. We then demonstrate that the texts are of different reading complexities using a readability formula — a simple statistical analysis of text used to determine how difficult it is to read. We then proceed with several analyses of the lexical links found in the texts. These examine both the different linking relationships, and their distribution in different books. Our principal finding is that the distribution of lexical linking relationships is independent of the type of text considered. A discussion of the results and their implications concludes the paper.

2. Selection of the Experimental Texts

There are several constraints on the selection of texts for the experiments. These are empirical, due to program requirements, and the limited nature of Roget's 1911 thesaurus (which is out of copyright hence readily available). The texts,

1. must be analysed within the constraints of the current implementation.
2. must be available electronically.
3. should be several thousand words in length.
4. should have demonstrably different complexities.

A range of texts of differing complexity was selected from those available on the Internet, or CD-ROM. These varied from children's books such as "Alice in Wonderland" to more challenging works such as Kant's "Critique of Pure Reason". The texts chosen are listed in table 1 below.

Table 1: Texts Selected

Title	***Author***	***Publication Date***
Alice's Adventures In Wonderland	Lewis Carroll	1867
Through The Looking Glass	Lewis Carroll	1867
Pride And Prejudice	Jane Austen	1813
Lectures on The Industrial Revolution in England	Arnold Toynbee	1884
Moby Dick	Herman Melville	1851
The Critique Of Pure Reason	Immanuel Kant[i]	1781

3. Reading Complexity of the Texts

The books used in these experiments were selected as representing a range of literary complexity. Books by Lewis Carroll are commonly read to junior school children, Austin and Melville are high school texts, whilst Kant and Toynbee are not usually encountered until University. Thus, we can expect intuitively that University level texts are harder to read than those aimed at school children. Nonetheless, some independent confirmation of their reading ease is desirable.

Readability is often measured by teachers to determine the suitability of books for pupils of different reading abilities. Readability formulae (e.g. Harrison 1980) aim to predict the level of a text's reading difficulty by calculating statistics, such as sentence length and mean syllables per word, from the text. They do not consider content, so need to be applied with caution.

Harrison (1980) describes ten readability measures, including the Flesch formula, and the Gunning FOG formula. Harrison (1980) reports a study by Lunzer and Gardner that shows that seven of the readability formulae are approximately correlated with pooled teachers' assessments of text reading levels.

Karlgren and Cutting (1994) showed that texts may be simply classified into fifteen different genres. They used the statistical technique of discriminant analysis on twenty parameters. These included sentence length, proportion of pronouns, average characters per word, and number of relative pronouns. They applied this method to classify the five hundred texts from the Brown corpus, which have been manually classified as belong to different genres. Karlgren and Cutting comment that readability measures work well to discriminate text types since they include the most salient features of their experiments including consider sentence length, word length, and characters per word.

The Flesch-Kincaid grade level measure computes readability based on the average number of syllables per word and the average number of words per sentence. It is a common metric that it is widely used. It is included in both Microsoft Word, and Corel's WordPerfect word processors, so it also has the advantage of convenience. The Flesch-Kincaid grade level was consequently calculated for the initial 1000 lines of the books in table 1. The 1000 line limit was chosen since this represents a reasonable subset of the book sufficient to capture its style. The results are shown in table 2 below.

[i] translated by J. M. D. Meiklejohn

Table 2 show that the books represent a range of reading complexity. They also demonstrate the internal consistency of the measure as two books of a similar style by the same author ("Looking Glass" and "Alice in Wonderland") have similar Grade levels.

We now move on to look analyse the data produced from the lexical chains identified the texts.

Table 2: Reading Complexity of the Texts

Book Title	Flesch-Kincaid Grade
Alice's Adventures In Wonderland	5.5
Through The Looking Glass	6.4
Pride And Prejudice	6.5
Moby Dick	7.8
Lectures on The Industrial Revolution in England	11.6
The Critique Of Pure Reason	12.0

4. Determination of the Lexical Cohesive Relationships

We used an algorithm based on those of Morris and Hirst (1991), and StOnge (1995) to identify the lexical cohesive relationships in the texts. Details are given in Ellman (forthcoming). Four relations were examined:

1. The links between identical words (hence ID)
2. Links between words that are not identical, but are member of the same Roget category (hence CAT)
3. Links between words that are members of the same group of categories in Roget, but not in the same category. (hence GRP)
4. Links through one level of internal thesaural pointers. (hence ONE)

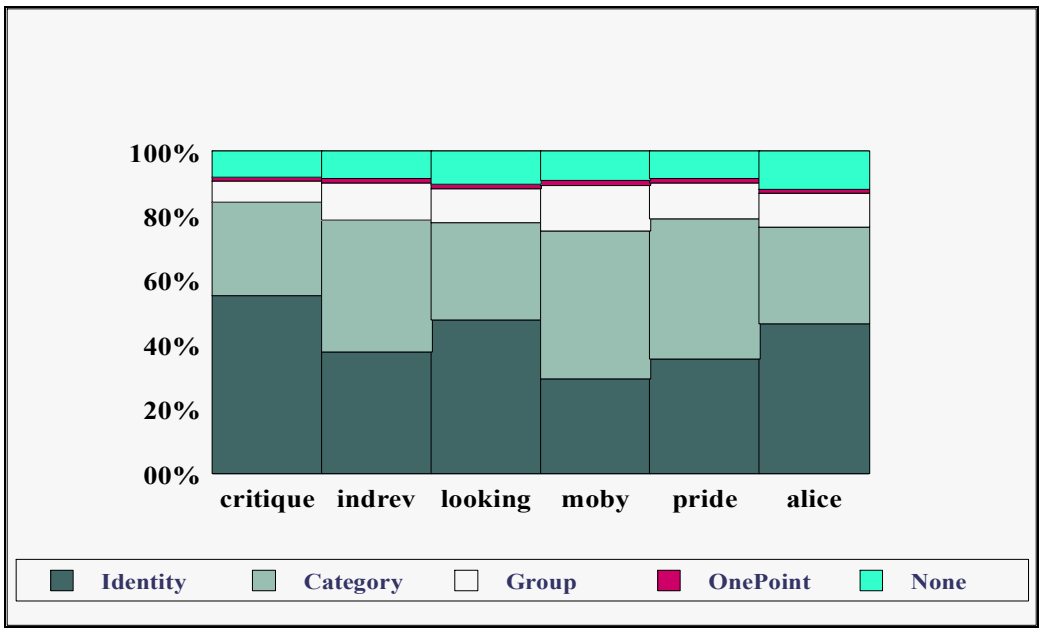
5. Analysis 1 : Link distribution between Documents.

This first analysis presents unprocessed sums of the link types. That is, all the lexical chains found in the documents were examined, and simple sums made of the types of lexical linking relationships found.

Our initial hypothesis was that there would more «weaker» linking relationships (such as GRP or ONE), since these can connect to a greater number of words than the identical word or same category relations. However, this was not the case.

Simple word identity (ID) is the most common lexical linking relationship found. Following that, we find Roget category entry (CAT) follows, then Roget group membership (GRP). The ONE relationship is relatively rare.

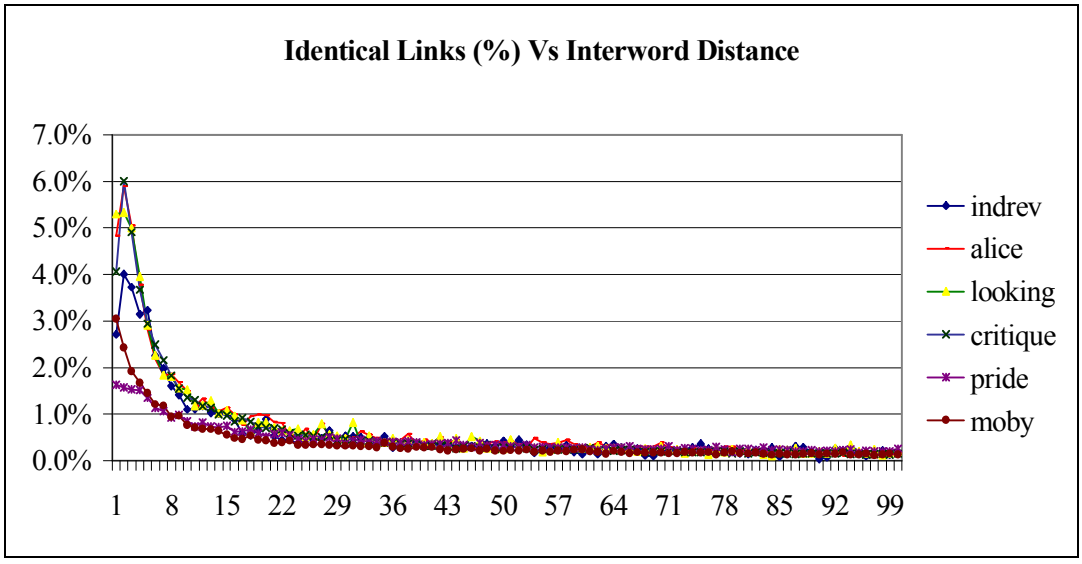
All the books in the experimental corpus show approximately the same total link distribution. Since the books represent increasingly complex texts, we have shown that the proportion of links of different types found in a text is broadly independent of the complexity of that text. We have also reason to question the value of the more complex thesaural relationships: The ONE level of indirection relation is sufficiently rare that one may question whether it is worth calculating.



Graph 1: Link Type Vs Book Title

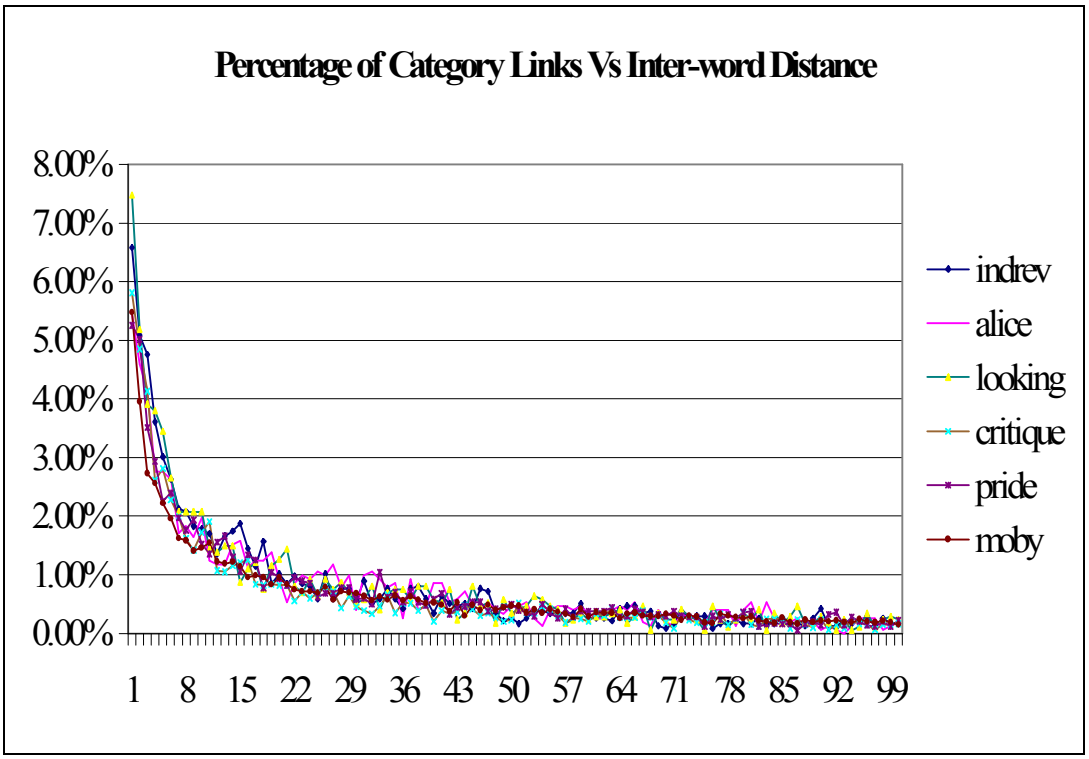
6. Analysis 2 : Link distribution change across different document types

Now we need to consider whether link distribution change across the different document types. This could arise if the threads of related words in the simple texts are shorter, hence making the text easier to read, or, alternately, denser text could have longer inter-word link distances. If this were to be the case the lexical chaining approach would not be a general tool, but would instead be some measure of text complexity.

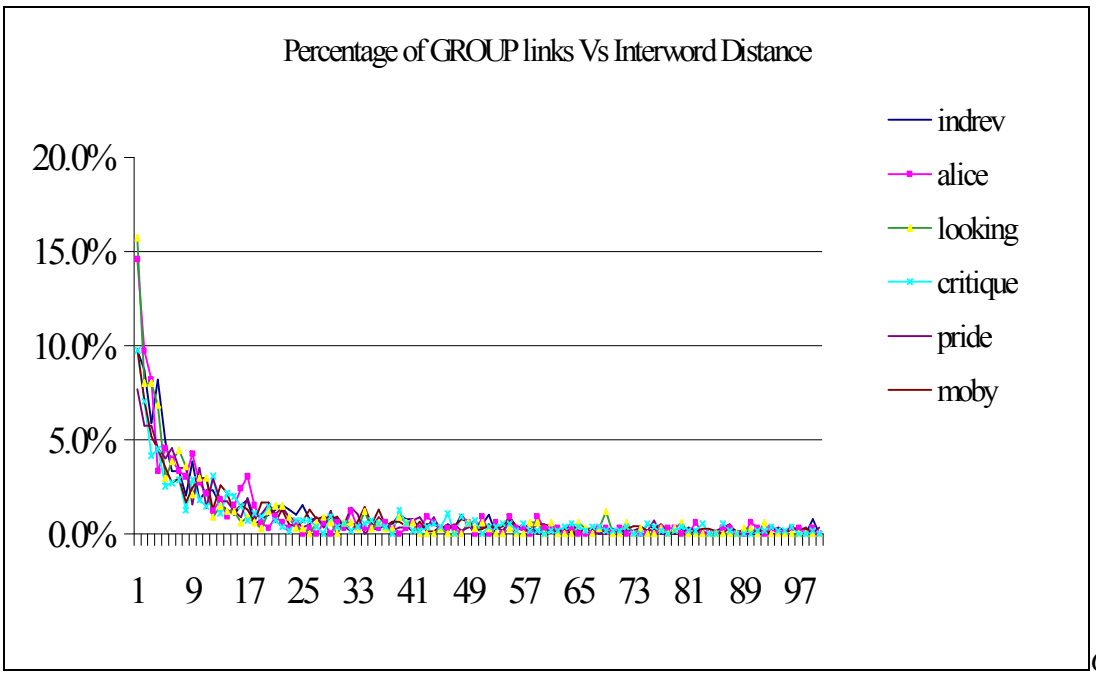


Graph 2: Identical Links (%) Vs Inter-word Distance

Comparative analysis is only possible if we compensate for the different lengths of the texts chosen. This is done by converting the number of links into percentages of the total links of that type. We can then plot the percentages of each link type that occur across the word window.



Graph 3: Percentage of Category Links Vs Inter-word Distance



Graph 4: Percentage of Group Links Vs Inter-word Distance

Graphs 2, 3, and 4 show the results of this analysis by link type[ii]. As can be seen, the percentage distributions are almost identical for all the texts, and for all the linkage types.

[ii] The ONE link type has been excluded since this does not occur frequently enough to generate consistent data.

This means that the type of text does not affect link creation in lexical chains. It also follows that the distance between words in a text is independent of the thesaural relationships sought between the words.

It can also be seen that Morris and Hirst had little justification for applying special status to the identical word relation, as they follow very similar distributions to the other thesaural links.

7. Conclusion

Lexical cohesion is a property of the words in a text. Relationships that link words have been termed lexical links. Links may be composed in to chains, and such lexical chains have great potential utility in text processing tasks, such as information retrieval, text similarity detection, or text summarisation.

A major concern is that types of lexical chains to be found in a text may depend on the style of that text. If this had been true, we would not have been able to base a measure of text similarity directly on lexical chains: it would have needed to be mediated by a determination of text genre

This has been disproved experimentally be analysing several book length texts. These were selected to be no more recent than Roget's 1911 thesaurus. This maximised the applicability of the lexical chaining algorithm. In addition to intuition, the books were shown to be of different reading difficulty by comparing them using the Flesch-Kincaid grade level readability measure.

An analysis of the distribution frequency of the lexical links found in the mini-corpus was strikingly similar for all the link types. This supports the hypothesis that text analysis measures based upon lexical cohesive links will be applicable to different styles of texts. Thus, the text similarity technique discussed in Ellman and Tait (1999) is capable in principle of determining the similarity of texts about the same subject, but written in different styles.

References

Beeferman D., Berger A., and Lafferty J. "A model of lexical attraction and repulsion. In Proceedings of the ACL-EACL '97 Joint Conference, Madrid, Spain

Bramer M., Macintosh A., and Coenen F (1999). "Research and Development in Intelligent Systems XVI" Springer-Verlag London UK ISBN 1-85233-231-X

Ellman J. 1998 "Using the Generic Document Profile to Cluster Similar Texts." In proceedings 1St UK Conf. On Computational Linguistics (CLUK).

Ellman J. and Tait T. (1999) "Roget's Thesaurus: An additional knowledge source for Textual CBR?" in Bramer et al. (1999).

Ellman J. (forthcoming) "Using Roget's Thesaurus to Determine the Similarity of Texts" PhD Thesis. School of Computing, Engineering and Technology, University of Sunderland UK.

Green S 1997 "Automatically generating Hypertext by Computing Semantic Similarity" University of Toronto PhD Thesis. Computing Systems Research Group Technical Report 366

Halliday, M.A.K. & Hasan, R.: 1976, "Cohesion in English", Longman, London.

Halliday, M.A.K. & Hasan, R.: 1989, "Language, context, and text". Oxford University Press, Oxford, UK.

Harrison C. "Readability in the Classroom" Cambridge University Press UK. ISBN 0 521 22712 7

Karlgren Jussi, 1999 "Stylistic Experiments" in Strzalkowski 1999

Karlgren, Jussi , and Cutting, Douglass 1994 "Recognizing Text Genres with Simple Metrics Using Discriminant Analysis" in proc. COLING 1994. (also http://xxx.lanl.gov/cmp-lg/9410008 accessed 30/6/99)

Lenz M. 1998 "Textual CBR and Information Retrieval - A Comparison." In proc. 6th German Workshop On Case-Based Reasoning. Berlin, March 6-8, 1998

Lesk M. "Automatic Sense Disambiguation using Machine Readable Dictionaries: How to tell a Pine Cone from and Ice Cream Cone

Miller G., Beckwith R., Felbaum C., Gross D., and Miller K. 1990 "Introduction to WordNet: An on-line lexical database" J. Lexicography 3(4) pp235-244

Morris J. and Hirst G. 1991 "Lexical Cohesion computed by thesaural relations as an indicator of the structure of text" Computational Linguistics Vol. 17 (1) pp 21-48

Okumura M. and Honda T "Word Sense Disambiguation and text segmentation based on lexical cohesion" Proc COLING 1994 vol. 2 pp 755-761

Sanderson M. 1994 "Word Sense Disambiguation and Information Retrieval" in Proc. ACM SIGIR 1994 pp 142-151

Smeaton, Alan 1999 "Using NLP or NLP Resources for Information Retrieval Tasks" in Strzalkowski 1999

Stairmand M 1996 "A Computational Analysis of Lexical Cohesion with Applications in Information Retrieval" PhD Thesis. UMIST Computational Linguistics Laboratory

Stokes A The reliability of Readability Formulae J. Of Research in Reading vol 1 1 1978 reprinted in Pugh, Lee, and Swann "Language and Language Use" Open University Press

St-Onge David 1995 "Detecting and Correcting Malapropisms with Lexical Chains" MSc Thesis University of Toronto. (available via WWW)

Strzalkowski Tomek 1999 "Natural Language Informational Retrieval" Kluwer Academic, Dordrecht NL. ISBN 0-7923-5685-3

Yarowsky 1992: Word Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. Proc COLING 1992 pp 454-460