The Linguistics of Readability: The Next Step for Word Processing

Neil Newbold
University of Surrey
Guildford
Surrey, GU2 7XH, UK
n.newbold@surrey.ac.uk

Lee Gillam
University of Surrey
Guildford
Surrey, GU2 7XH, UK
l.gillam@surrey.ac.uk

Abstract

In this paper, we present a new approach to writing tools that extends beyond the rudimentary spelling and grammar checking to the content of the writing itself. Linguistic methods have long been used to detect familiar lexical patterns in the text to aid automatic summarization and translation of documents. We apply these methods to determine the quality of the text and implement new techniques for measuring readability and providing feedback to authors on how to improve the quality of their documents. We take an extended view of readability that considers text cohesion, propositional density, and word familiarity. We provide simple feedback to the user detailing the most and least readable sentences, the sentences most densely packed with information and the most cohesive words in their document. Commonly used verbose words and phrases in the text, as identified by The Plain English Campaign, can be replaced with user-selected replacements. Our techniques were implemented as a free download extension to the Open Office word processor generating 6,500 downloads to date.

1 Introduction

Spell and grammar checking have become inherent tools in many modern word processors even if their results are not always deemed appropriate. Work on writing tools has largely focused on improving these services, with superior grammar checkers being the emphasis of this work. However, there has been little effort on providing a deeper analysis of the text, such as covering its semantic content and its potential success in conveying the authors intended message to the reader. Research on readability aimed to provide an indication of the proportion of the population could understand the text but has been limited to simple checks of word and sentence length providing only some degree of feedback on where and why text is difficult to understand. Writing tools such as ‘Stylewriter’ scores documents based on average sentence length, number of passive verbs and overall style. The style analysis uses an indexes check for a wide variety of common editorial issues like jargon, hyphenation, sexist writing, clichés, grammar, redundancies and troublesome words which are either abstract, complex, misused or overused. However, their approach is based on a simple lookup of common writing patterns with no analysis of overall message clarity. More robust tools such as ‘Coh-Metrix’ (Graesser et al., 2004) deliver a substantial analysis but can leave casual users confused with the quantity of numerical data produced.

In this paper, we discuss how linguistic techniques have been deployed to measure largely ignored aspects of the text, which can benefit authors when writing texts. We use automatic summarization techniques to measure how cohesive or consistent the text is and parts of speech patterns to identify multi-word expressions, which indicate portions of text densely, packed with information. We also deploy corpus linguistics methods to measure the familiarity of words in everyday use. These techniques expand upon readability research to provide a series of tools for authors giving pointers to where their documents might confuse their intended audience.

Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing, pages 65–72,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

2 Background

In principle, readability measures identify some proportion of the population who could comfortably read a text. Historically, readability research has focused primarily on producing a numeric evaluation of style of writing to associate textual content to a particular rating or the level of education of readers. Readability research largely traces its origins to an initial study by Kitson (1921) who demonstrated tangible differences in sentence lengths and word lengths, measured in syllables, between two newspapers and two magazines. Kitson's work led variously to the development of readability metrics, many of which are available in certain software applications. Further discussion of these formulae can be found elsewhere (Dubay, 2004). More recent considerations of readability account for reader factors, which consider certain abilities of the reader, and text factors, which consider the formulation of the text (Oakland and Lane, 2004). Reader factors include the person's ability to read fluently, level of prior subject knowledge, lexical knowledge or familiarity with the language, and motivation and engagement. Text factors account to some extent for current readability metrics, but also cover considerations of syntax, lexical selection, idea density, and cognitive load. Oakland and Lane's view of readability suggest that it may be possible to generically measure the difficulty of text as an artifact, but that "text difficulty" necessitates consideration of each reader. Our work elaborates that of Oakland and Lane in identifying difficulties in the apparently neat separation of the factors. In this section, we propose a new framework for readability that builds on Oakland and Lane by making consideration of the relationship between text, reader, and author. We explore, subsequently, how word processors might use such a framework to help authors get across their intended messages.

2.1 Matching Text to Readers

In writing a document, an author has to be mindful of the needs of his anticipated audience, particularly if they are to continue reading. There must be some correlation across three principal aspects of a text: the nature and extent of its subject matter, its use of language, and its logical or narrative structure. The audience can be defined by their degree of interest in the subject, how much they already know about it, their reading ability, and their general intelligence. If the author needs to learn more about a set of potential readers, standardized tests are available to measure levels of intelligence and reading skill, while interest and prior knowledge can be assessed by ad hoc surveys.

Two kinds of measure are suited to appraising text structure: logical coherence and propositional density. By logical coherence, we mean the extent to which one statement is ordered according to a chain of reasoning, a sequence or chain of events, a hierarchy or a classificatory system. By propositional density we mean the closeness, measured by intervening words, between one crucial idea and the next. The less coherently ordered are it's the ideas and the greater their density, the larger the cognitive load on the reader.

If characteristics of the audience have been ascertained, the author must ensure that what he is writing is generally suitable for them. A readability formula will produce a quick check on a given text for an author, and comparisons have been made amongst measures to correlate with specific human performances over largely disjoint sets of texts. For our current considerations, we are interested in providing more useful feedback to the author that a single numerical value. A readability analysis should be able to provide hints to the author on how to improve their text. However, this is not to say that existing measures are adequate, we propose other elements of text that can be measured instead of, or in addition to those examined by the currently established readability formulae. Our new framework for readability, describing the factors to be considered is presented in Fig. 1. The matches needed for easy reading describes how an author can match their text to their target audience. In the remainder of this section, we elaborate these factors.

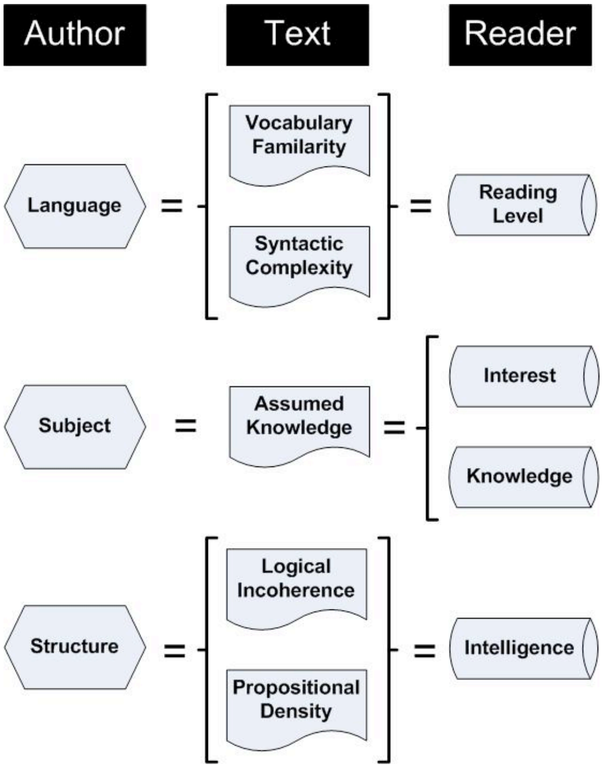

Figure 1: Matches needed for easy reading

2.2 Language

When matching text to reader, the author needs to consider the level of language and style of writing. This can be described as the vocabulary familiarity and syntactic complexity of the writing. These aspects are generally measured by the existing readability formulas as word length and sentence length. Readability metrics generally determine the difficulty of a word by counting characters or syllables. However, Oakland and Lane suggest that word difficulty can be determined by examining whether the word is challenging, unusual or technical, and cite word familiarity as more effective means of measuring word difficulty. The process by which readers develop word familiarity is through their language acquisition and the development of their language capability. Frequency plays an important role in building knowledge of a language so that it is sufficient to understand its written content. Diessel (2007) showed that linguistic expressions stored in a person's memory are reinforced by frequency so that the language user expects a particular word or word category to appear with a linguistic expression. These linguistic expectations help comprehension.

Frequency was also found to be fundamental in reading fluency as words are only analyzed when they cannot be read from memory as sight words. A limited knowledge of words affects reading fluency as readers are likely to dwell over unfamiliar words or grammatical constructions. This impedes the reader's ability to construct an ongoing interpretation of the text. The reading fluency of the reader is dependent on their familiarity with language. When readers find text populated with unfamiliar words it becomes harder for them to read. This is especially prevalent in scientific or technical documents where anyone unfamiliar with the terminology would find the document hard to understand. The terminological nature of specialized documents means that terms will appear with disproportionate frequently throughout the documents in contrast to what one would expect to encounter in everyday language. Terminology extraction techniques exploit this relationship to identify terms. We adapt this method by contrasting word frequency within documents with familiarity in general language. We determine the difficulty of a word by its familiarity.

Vocabulary does not tend to exist in isolation. The vocabulary may be well-defined, yet included in overly verbose sentences. Consider these two sentences:

1. "We endeavor to maintain the spinning of all the plates."
2. "We try to keep all the plates spinning."

The first sentence uses passive voice, the second uses active voice. Writing guidelines, such as those presented by the Plain English Campaign (1979), often recommend active voice wherever possible. Active voice uses fewer words and helps readers build a mental representation of the text. Existing readability formulae consider that long and complex sentences can confuse the reader and whilst we support this view, we consider that each individual sentence should be scored to allow the author to identify the particularly troublesome sections of their text. When matching text to reader, syntactic complexity should be examined not just for the entire document but whether each sentence is appropriate for the reading level.

2.3 Subject

To learn from text, a reader needs to associate the new information to their existing knowledge. This task can be helped by the reader's interest level. Kintsch et al. (1975) showed that we find stories easier to remember than technical texts because they are about human goals and actions, something to which we can all generally relate. Scientific and technical texts require specific knowledge that is often uncommon, making the texts impenetrable to those outside the domain. This suggests that readability is not merely an artifact of text with different readers having contrasting views of difficulty on the same piece of text. Familiarity with certain words depends on experience: a difficult word for a novice is not always the same as a difficult word for an expert. Reader characteristics such as motivation and knowledge may amplify or negate problems with difficult text. When matching text to reader, the author needs to consider the target audience and the extent of their knowledge.

Many readability metrics do not make distinctions based on the background knowledge of the reader. As discussed in relation to vocabulary, word familiarity can give a better indication of word difficulty than word length. A longer word may only be difficult for a particular reader if unfamiliar, and certain shorter words may even be more difficult to understand. Consider a general reader confronted in text discussing a 'muon': This short term would be rated as simple by current readability formulae. However, a majority of people would be unfamiliar with this term, and only physicists are likely to know more about the term, its definition, and related items. One way to measure background knowledge would be to consider the extent of use of known terms in the text with direct consideration of previous documents within the reader's experience.

Entin and Klare (1985) showed that more readable text is beneficial for those with less knowledge and interest. In their study, students were presented with written material below their reading level. When the reader's interest was high, text below their grade level did not improve comprehension. However, when the reader's interest was low their comprehension was improved by simpler text. This suggests that more readable text improves comprehension for those less interested in the subject matter. We consider the need to capture and analyze the user's experience with prior documents as a proxy for reader knowledge and motivation. Given a reading history for a user, we might next build their personalized vocabulary with frequency information, and therefore measure familiarity with words on an individual basis. In the same way that an expert is familiar with the terminology of their subject, we can reflect the background knowledge required by a reader to interpret the text correctly. When matching text to reader, word difficulty should be measured, if the information is available, on an individual basis.

2.4 Structure

Well-written text requires a structure that readers can readily use to find the information they need and to understand it correctly. Text can become confusing when information is inappropriately presented. Most sentences, when taken out of context, can become multiply ambiguous. When we read text, we build a collection of the concepts described within it. We identify these concepts with words and phrases using pragmatic, semantic, and syntactic features. We build certain interpretations with these blocks of words that tend not to combine randomly or freely, but rather they keep preferred company (Firth, 1957). These collocations are evidence of preference for certain friends, and these friends may be kept at certain distances. For example, words impose restrictions over synonyms, excluding some from their group of friends so that 'strong tea' may be acceptable, but 'powerful tea' may not. A reader unfamiliar with such constructions might not understand the precise meanings or variations. In addition, individual words may not be particularly difficult but their combination may produce different meanings to the component words. Collocation statistics may indicate compound nouns with specialized meaning but increased likelihood of misinterpretation. Consider, for example, 'glass crack growth rate': each word should be relatively easy to understand, but interpretations due to bracketing (Pustejovsky et al., 1994) might lead to interpretations of a 'crack growth rate' made of 'glass', and an unpacking of semantics may be useful in removing ambiguities due to bracketing.

Most researchers agree that collocations are sequences of words that co-occur more often than by chance, with certain assumptions of randomness, and can be found using statistical measures of association. Some linguists consider collocations are the building blocks of language, with the whole collocation being stronger than the sum of its parts. They describe collocations as lexical items that represent uniquely identifiable concepts or semantic units. Smadja (1993) elaborated criteria for a collocation, describing them as recurring and cohesive domain-dependent lexical structures such as 'stock market' and 'interest rate', and suggested how components can imply collocations, for example 'United' produces an expectation of 'Kingdom', 'Nations', or 'States'. When frequently combined linguistic expressions develop into a processing unit, many of the linguistic elements are ignored and the whole chunk is compressed and treated as one semantic unit. These units often develop into terms with multiword units representing singular concepts. This relates back to the assumed knowledge of the reader. However, for readers unfamiliar with the terms, we have identified two methods called 'Propositional Density' and 'Lexical Incoherence' for processing semantic units.

When a significant amount of information is conveyed in a relatively small amount of text, the reader can become confused. We identify this problem as 'Propositional Density'. Although long collocations form semantic units that reduce conceptual complexity, problems occur when numerous semantic units are described within a short space of each other causing the reader to make numerous inferences. The number of ideas expressed in the text contributes to the work required of the reader to interpret the text correctly. Propositional density may be measurable by examining the quantity of objects within short distances of each other. These objects can be labeled with single nouns or multi-word expressions. By measuring the number of unique semantic units, we can approximate the workload required for processing or interpreting the text correctly.

The second problem with text structure is called 'Lexical Incoherence' and occurs when writers present new information to the reader without making clear its relationship to previous information. The writer assumes that they have provided enough information to allow readers to follow their arguments logically. Repetition of concepts, terms, and other referents provides a structure for the reader to connect with. It is through this repetition that a series of links can be made between the sentences. There is a relationship here to work on lexical cohesion (Hoey, 1991). If a large number of new, seemingly unrelated ideas are being introduced, low cohesion would be expected and measurable. Efficiency can be increased here by using synonyms. Semantic units can be referred to by a number of different labels and by identifying these different labels we can more accurately find the prominent ideas in the text.

3 Open Office Readability Report

To implement our new techniques for measuring readability, we used OpenOffice.org 3, which is the leading open-source office software suite. As it can be downloaded and used free of charge, it has an already established user base and allows third-party developers to write extensions for their applications. These extensions are made available to download for any OpenOffice.org user. We created the readability report extension for 'Writer', the open office word processor, to implement our readability techniques. The extension generates 5 separate components devised from our framework for readability incorporating the matches for easy reading. The components analyze the text factors in the framework to provide an indication of the corresponding reader factor. The author can use this information to help match their text to their audience. Each author and reader element is addressed by a component as follows:

- Language -> Reading Level
    - Weirdness Measure
    - SimpleText SmartTags
- Subject -> Interest and Knowledge
    - Not yet implemented
- Structure -> Intelligence
    - Propositional Density
    - Lexical Coherence

The two language components address both the text features of vocabulary familiarity and syntactic complexity. We have yet to implement a component to assess the subject of the text. The separate components, which consist of either a generated report or text annotation through SmartTags, are detailed in the remainder of this section.

3.1 Weirdness Measure

The first generated report uses our new readability formula based on word frequency. Unlike the established readability formulas, our measure can be applied to an individual sentence, allowing the report to highlight the most and least readable sentences in the document. We use frequency information from the 100 million word tokens of the British National Corpus (BNC) to act as a reference corpus. The frequency counts for each word along with the number of words in the sentence are used to determine the sentence readability. The score for the document can then be ascertained as an average value for each sentence. We use log and other arbitrary values to bring the final number into a similar range as the other readability formulas.

$$2.1 \times \ln(n_{SL} \sum \sqrt{\frac{f_{SL} n_{GL}}{(1 + f_{GL}) n_{SL}}}) \tag{1}$$

Eqn. 1 sentence-based familiarity
f is word frequency, n is word count at sentence level (SL) or corpus level (GL).

Based on research by Stuart et al., (2004) concerning the frequency of use of apostrophes by children, contractions such as "n't" and "'s" were considered as separate words with their difficulty determined by their frequency count as per any other word. This method for analyzing contractions generated more effective results from the BNC.

In addition, to the weirdness measure the report provided the scores from the established readability formulas, 'Flesch Easy Reading', 'Flesch Kincaid', 'FOG', 'SMOG' and 'ARI'. To help authors understand the significance of the readability values, a series of ratings were provided for each measure (inc. Weirdness), which grade a document as either 'Simple', Easy', 'Good', 'Challenging' or 'Difficult' using a series of threshold values.

3.2 SimpleText SmartTags

SmartTags were developed in Open Office to highlight sections of documents and add contextual information. The readability report extension uses SmartTags to highlight difficult words and phrases in the text, as identified by http://www.plainenglish.co.uk/. The 'SimpleText SmartTags' provide suitable alternatives for these phrases which can be inserted automatically into the text. The user can click on the SmartTags and select a possible alternative from a list of suitable replacements. These SmartTags help authors avoid using common, verbose expressions that hinder the clarity of their writing. Gillam and Newbold (2007) showed that plain English substitutions can lower readability scores.

3.3 Propositional Density

To measure the structure of the text, we use an analysis of propositional density. This report analyses how many concepts and ideas are referred too in the text and was entitled the 'Brain Overload Report' to make it more accessible to the users. An expert in a particular subject will often use specific terms and jargon resulting in too much information being presented to the reader within a short space. This can lead to learners become fatigued and confused. The report measured the amount of single and compound nouns in comparison to the length of the sentence in which they occurred. Sentences with contained a large amount of 'glue' words, such as 'the', 'at', etc. would score lower than sentences loaded with multi-word expressions. The score for the document is determined as an average value for each sentence. For user convenience, we use some arbitrary values to bring the final number into a similar range as other readability formulas.

$$2n \frac{u + n}{3(n - c)} \tag{2}$$

Eqn. 2 sentence-based propositional density
u is the number of semantic units, n is sentence length, c is the number of collocated words.

Whilst this report was devised to measure the structure of the text, there is an also an element of subject which determines the assumed knowledge of the reader. Scientific and technical texts often require specific knowledge represented in the text through frequent use of terminology. The single and multi-word terms increase the propositional density of the document indicating that the text will be difficult for novices in the subject matter. Texts intended for a general audience should score low on propositional density. As with the previous report, the resulting score is graded as either 'General', 'Introductory', 'Scholarly', 'Technical' or 'Specialized' to help authors understand the impact their text will have on their intended audience.

For further feedback, the most frequent multiword expressions are listed to show authors which expressions are contributing the most to their score. Each expression is unpacked into its component expressions and a frequency count throughout the document is taken for each. The most frequent component expression is then used as a basis for unpacking the full expression. For example, 'current account balance' will be unpacked into either 'balance of current account' or 'account balance which is current', depending on which of the component expressions, 'current account' and 'account balance' are more frequent. If a suitable component expression is found, the full unpacked expression is suggested to the user as a possible way of rewriting the collocation. The separating glue words are selected depending on the Part-Of-Speech tagging of the concluding phrase in the rewritten expression.

3.4 Cohesion Measure

The 'Cohesion Report' uses techniques for automatic summarization to measure how easy a document is to follow. It identifies the lexical words in each sentence and uses them to recognize sentence bonds. Hoey (1991) described a sentence bond as two sentences sharing 3 or more lexical words. The score for a document is determined by the number of sentence bonds against the total number of possible sentence bonds.

$$\frac{b}{s(s-1)} \qquad (3)$$

Eqn. 3 document-based lexical coherence
b is the number of sentence bonds, s is the number of sentences.

The sentence with the highest number of bonds is highlighted in the report as the most representative of the document. For further feedback, the report shows the words that are the strongest themes in the text. These are lexical words that were used the most often to create sentence bonds. By increasing the references to these themes the author can improve the cohesion of their text. Authors need to pay particular attention to sentences with no sentence bonds are these are adding nothing to the coherence of the text. These can be seen by using the detailed report described in the next section. The cohesion measure is primarily useful for documents about a specific subject; fictional writing will often score low for cohesion. As with previous reports, a document will be graded as either 'Creative', 'Digressing', 'Consistent', 'Coherent', or 'Fluent'.

3.5 Detailed Report

An option is provided for a detailed report that allows authors to view the readability score of each sentence in their document. The report is displayed in a spreadsheet and shows the results of each of the established readability measures and the new scores discussed in this paper. The spreadsheet can be used to identify the most troublesome sentence in the document. This is particularly useful for examining the results of the cohesion measure, as sentences that are not adding to the cohesion of a document can be easily identified.

4 Conclusion

The weirdness measure correlates with the other readability formulas that have been shown to indicate the required reading age of the text, when analyzing a large range of texts. Our results show that frequency is a good indicator of word difficulty. Table 1 shows a sample of texts, ordered by increasing difficulty, ranging from children's books to technical reports and the correlation of the weirdness measure to the established formulas. For sentences, containing relatively long but commonly used words such as 'information' and 'business', the score calculates more probable figures than the established readability formulas. Certain children's books (for example "Jabberwocky") contained made-up or nonsense words which caused the measure to the rate the texts as difficult. It should be noted that the other readability formulas rated these texts as simple. We consider that these types of text would be confusing to non-native speakers of the language, with the effect of these words, which are unique to the document, being the same as terminology.

Text	Weird	Kincaid	FOG	SMOG	ARI
Lucky	8.54	3.80	5.26	6.74	2.75
The Absolutely True Diary	9.40	4.62	6.31	7.05	3.33
Coraline	9.41	5.08	7.24	8.05	4.41
Associated Press, Fed Revises	11.78	10.92	12.50	11.96	11.20
Bloomberg, U.S. Leading Indicators	11.76	11.28	13.33	12.60	11.51
USA Today, Greenspan predicts	12.07	11.36	13.25	11.21	12.27
Greenspan, to congressional committee 2005	12.75	14.32	16.29	14.60	14.55
Greenspan, speech 2005	12.52	15.69	17.80	15.70	16.18
Bernanke, speech 2008	13.29	16.28	17.97	15.58	17.72
Bernanke, report to congress	13.24	16.60	18.80	16.31	17.80
Correlation		**0.98**	**0.98**	**0.96**	**0.99**

Table 1: Correlation of weirdness measure with established readability formulas

Currently, we have not implemented a means to measure the user's experience with prior documents as a proxy for reader knowledge and motivation. In future, we would build a personalized vocabulary for each user with frequency information, and therefore measure familiarity with words on an individual basis. At present the Open Office extension is more useful for writing general texts as opposed to specialized or technical documents. Certain elements such as the weirdness measure and the SimpleText SmartTags become less useful with more expert texts and the prevailing use of terminology. Other aspects such as propositional density and lexical coherence are of less use when analyzing children's books. This style of writing can score high for propositional density due to extravagant character names (e.g., "The Mad Hatter" and "Cheshire Cat") increasing the number of compound nouns. Lexical cohesion is also low for any fictional writing.

We are looking at improving the lexical cohesion measure with the consideration of synonyms. Semantic units can be referred to by a number of different labels and by identifying these synonyms; we can more accurately identify the prominent ideas in the text. It is the repetition of terms and their synonyms, along with other referents that provide a structure for the reader to connect with.

The extension was made available on the Open Office website in July 2009. In six months the extension had received over 6,500 downloads indicating, along with positive user feedback that demand for word processors to go beyond simple spelling and grammar checking of text and provide more feedback is considerable.

References

H. Diessel. 2007. Frequency effects in language acquisition language use, and diachronic change. *New Ideas in Psychology*, 25(2):108–127.

W. H. DuBay. 2004. *The Principles of Readability*. Costa Mesa, CA: Impact Information.

E. B. Entin, G. R. Klare. 1985. Relationships of measures of interest, prior knowledge, and readability comprehension of expository passages. *Advances in reading/language research*, 3: 9–38.

J. R. Firth. 1957. *Papers in Linguistics: 1934-1951*. London, Oxford University Press.

L. Gillam, and N. Newbold. 2007. *Quality assessment. Deliverable 1.3 of EU eContent project LIRICS*, URL:http://lirics.loria.fr/doc_pub/T1.3Deliverable.final.2.pdf, last accessed 28 February 2010.

A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai. 2004. Coh-Metrix: *Analysis of text on cohesion and language. Behavior Research Methods, Instruments, and Computers*, 36:193–202.

M. Hoey. 1991. *Patterns of Lexis in Text*. Oxford, OUP.

W. Kintsch, E. Kozminsky, W. J. Streby, G. McKoon, and J.M. Keenan. 1975. Comprehension and recall as a function of content variables. *Journal of Verbal Learning and Verbal Behavior*, 14:196–214.

H. D. Kitson. 1921. *The mind of the buyer*. New York, Macmillan.

T. Oakland and H. B. Lane. 2004. Language, reading, and readability formulas: Implications for developing and adapting tests. *International Journal of Testing*, 4(3):239–252.

The Plain English Campaign. Established 1979. URL:http://www.plainenglish.co.uk/, last accessed 28 February 2010.

J. Pustejovsky, S. Bergler, P. Anick. 1994. Lexical semantic techniques for corpus analysis. *Computational Linguistics*, 19(2):331–358.

F. Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–178.

M. Stuart, M. Dixon, and J. Masterson. 2004. Use of apostrophes by six to nine year old children. *Education Psychology*, 24(3): 251–261.