Populating a framework for Readability Analysis
Word frequency = word difficulty

Neil Newbold and Lee Gillam
University of Surrey
n.newbold@surrey.ac.uk and l.gillam@surrey.ac.uk

Abstract

This paper discusses a computational approach to readability that is expected to lead eventually towards a new and configurable metric for text readability. Our research involves the elaboration, implementation and evaluation of an 8-part framework that requires consideration of both textual and cognitive factors such as language, vocabulary, background knowledge, motivation and cognitive load. We will discuss our work to date that has examined the limitations of current measures of readability and made consideration for how the wider textual and cognitive phenomena may be accounted for. This includes techniques for statistical and linguistic approaches to terminology extraction, approaches to lexical and grammatical simplification and considerations of plain language. Our interest in readability is motivated towards making semantic content more readily available and, as a consequence, improving quality of documents, There are already indications that this work will contribute to a British Standard for readability.

1. Introduction

Text is used worldwide to convey information. With the ever-expanding web, written communication has become more important than ever been before. The burden on users in filtering, understanding and rapidly processing large volumes of written communication is mounting. For users to achieve the most effective understanding of content in a short time, the need for clear and concise writing gains ever greater importance. Yet the advent of social media, blogs, instant messaging and SMS texts has put emphasis in the opposite direction, on speed of communication at the detriment of clarity. The use of readily understandable and consistent language is essential for enabling readers to understand written text: it affects their ability to comprehend and assimilate what a writer thinks they are conveying. However well-written text is not just advantageous to human readers, automatic text processing systems can also benefit from text which is easy to interpret and free from ambiguity. For natural language processing systems attempting to interpret semantic content from text, convoluted and unnecessarily verbose information can cause considerable problems. Often such systems require extensive manual training and/or "eyeballing" of texts, and their effectiveness depends on this. The aim of this work is to explore the potential for automation improving the readability of text to make its semantic content more accessible. This process involves a variety of corpus linguistics techniques, typically used in isolation, which can be combined as a means to achieve this. Such efforts have multiple potential outcomes: (i) improved human understanding and transference of concepts between author and reader; (ii) improved machine processing, such as faster and more accurate building of text corpora; (iii) improved capacity for filtering, using document quality as an additional measure, to avoid opaque texts.

In section 2 we provide an overview of current readability methodologies and practices and how these principles have been applied in the past. We explore a variety of techniques for improving text by examining factors concerning both the actual text and the reader. Section 3 focuses on how to analyse and improve text for human readers using the domain of quality assurance. There has been little work on concerning quality assurance and textual quality. Terminology, jargon and ambiguity compound the clarity of text and an over-reliance on ill-defined or author-invented terminology leads to the word "jargon" being applied in a pejorative sense. Important messages in the text, intended to be conveyed by the author or expert annotator, need to be highlighted and not concealed. Authors of international standards, such as ISO, demand that written work be precise and comprehensible. However, there is only a small amount of written "guidance" on how to do so, and where it does exist it is easily ignored. In this section we automatically analyse draft ISO documents and provide recommendations for authors on how to improve their writing. The analysis uses of a combination of terminology extraction, linguistic techniques and readability formula to identify potential problems for readers of the documents which the author can choose to incorporate. Once any amendments have been made the document can be re-analysed iteratively to the authors' satisfaction. Section 4 discusses our new readability formula using frequency counts from text corpora as a means to determine word difficulty. The frequency list can be substituted for any list the authors seems suitable for any subject or language. Section 5 concludes the paper and makes some considerations for future work.

2. Background

Readability is a measure of how easy text is to understand, indicating how wide an audience it will reach. It is often confused with legibility and the W3C consortium on accessibility focuses on presentation issues such as typeface, text size, layout and colours in web pages but does not consider the content of the text. It is the message, not the medium that is of concern to readability researchers. Klare (1963) defines readability as "the ease of understanding or comprehension due to the style of writing". According to this definition, the abilities of the reader are not an important factor for readability. Dale and Chall (1949) incorporate reader characteristics such as reading fluency and motivation, and describe the success of readability as "the extent to which they (readers) understand it, read it at an optimal speed, and find it interesting". For us, the most comprehensive consideration of readability to date is presented by Oakland and Lane (2004). They have emphasized that readability entails two considerations: reader factors and text factors. Both contribute in a variety of ways to text difficulty. These factors are shown in figure 1. Reader factors concern the person's ability to read fluently, whether they have sufficient background knowledge in the subject, their lexical knowledge or familiarity with the language, and whether they are suitably motivated and engaged in the subject matter. Text factors cover syntax, lexical selection, idea density and cognitive load - the effort required by the reader to correctly interpret the text.

All of these factors have an effect on the reader's ability to understand text. By exploring each of Oakland and Lane's factors contributing to text difficulty, we hope to provide a more conclusive analysis of readability. We aim to integrate the different approaches to readability and provide a framework which is stronger than its components. This paper will demonstrate and evaluate a computational model for readability as an integration of Oakland and Lane's framework with natural language processing techniques such as terminology extraction and lexical cohesion. This work will lead to new tools and measures for determining readability which can be used by domain experts and novices alike.

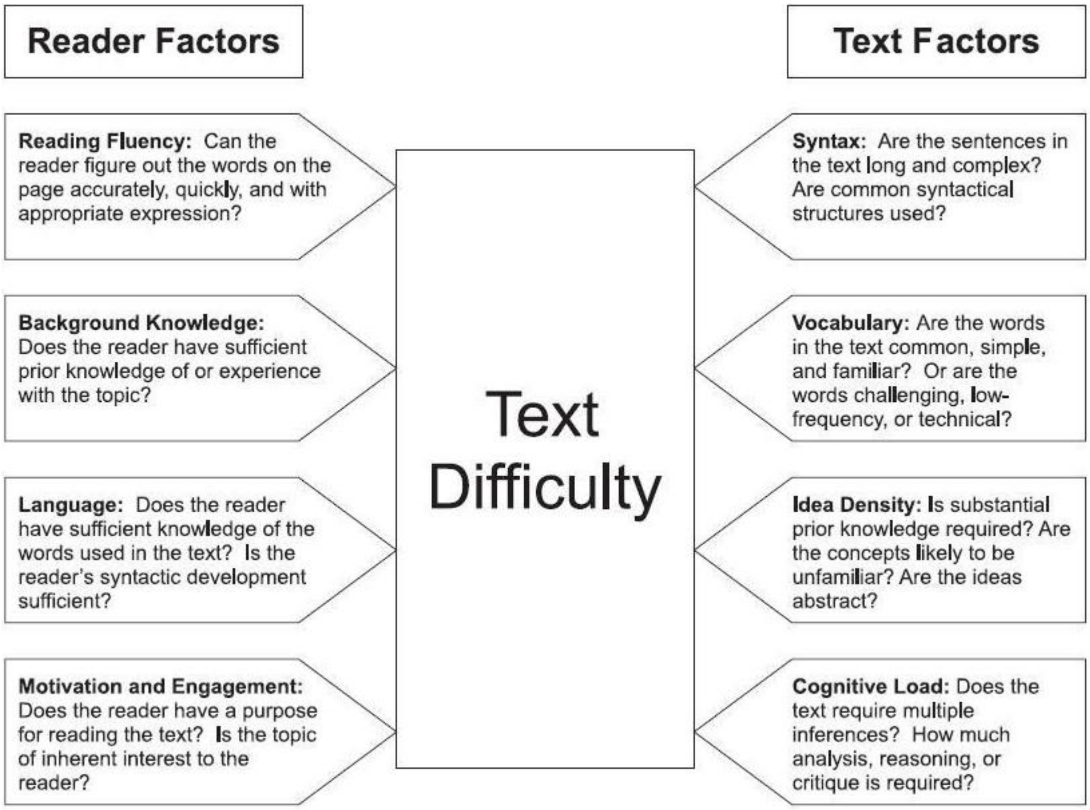

Figure 1. Oakland and Lane's factors contributing to text difficulty.

2.1 Text factors

The relationship between text and the reader has long been the focus of readability research. An early study by the psychologist Kitson (1921) showed how readers of various magazines and newspapers differed from each other. He analysed two newspapers and two magazines by examining 5000 consecutive words and 8000 consecutive sentences in the four publications. He found that the sentence length and word length measured in syllables were shorter in one newspaper and one magazine than they were in their counterparts. Kitson proposed that these variables accounted for the differences in readership. Kitson's work led to the development of other metrics for readability based on considerations of sentence length and word length, and in some cases as a function of the number of syllables. Many of these have proven to be popular in modern office software applications. One of the most common methods for readability measurement is the Kincaid Formula developed by Kincaid, Fishburne, Rogers and Chissom (1975). The result from this measure is the U.S. school grade required to easily understand the text on the first reading. The ideal grade level is or 8 (ages 13 and 14) with results of more than 12 running a serious risk of not being understood. The American Navy use this method to judge the readability of their technical manuals and many government agencies use readability measures to ensure documents or forms meet their specific readability levels. A test devised by Flesch (1948), called the Flesch Index or the Flesch Easy Reading Formula has become a U.S. governmental standard. The index is between 0 and 100 with a higher score indicating the easiness in understanding the document. The average English document usually achieves between 60 and 70. The measure is noted for being significantly more sensitive to long words than the Kincaid Formula.

Another widely adopted readability formula is the Fog Index, devised by Gunning (1952) who claimed that the index does not indicate how good the actual writing is but rather how easy it is to understand. Gunning believed that measuring the amount of what he called “mental fog” in a piece of writing was not an exact science and that the score produced by his test was only a rough indication of text difficulty. Like the Kincaid formula, the Fog Index produces the grade level needed to understand the text with the "ideal" Fog Index being around 7 or 8. The formula incorporates a count of complex words which are defined as words with three or more syllables, discounting common suffixes such as -es, -ed, or -ing. Compound words and proper nouns are also not counted as complex words. McLaughlin (1969) used the complex word count as an element of SMOG (Simple Measure of Gobbledygook) which was named as a tribute to Gunning’s Fog Index. SMOG is considered more accurate and easier to calculate than Gunning’s formula and like the Fog Index produces the Grade level as a result. However, one of the problems with both the Fog and SMOG measures is that not all multisyllabic words are difficult to understand. For example, "spontaneous" is generally not considered to be a difficult word, despite the fact it has four syllables. Smith and Senter (1967) devised a formula called the Automated Readability Index (ARI) which relied on characters per word, instead of the usual syllables per word. The intention was to produce a more accurate automated readability measure as computer programs can count the number of characters more accurately than the number of syllables. However, despite its ease of implementation, the accuracy of ARI is much disputed. Table 1 shows the formulae discussed and the elements on which they rely.

	Flesch	**Kincaid**	**Fog**	**SMOG**	**ARI**
Sentence length	✓	✓	✓	✓	✓
Characters/word					✓
Syllables/word	✓	✓			
Complex words count (more than three syllables)			✓	✓	
Scale	0-100	US Grade level	US Grade level	US Grade level	US Grade level
Ideal outcome	100	7-8 (Ages 13-14)	7-8 (Ages 13-14)	7-8 (Ages 13-14)	7-8 (Ages 13-14)

Table 1. Features of readability metrics.

There are two main types of application for readability metrics:

1. for educators in selecting appropriate material for the target audience’s reading ability or to determine whether feedback comments provided to students will improve learning outcomes (Williams and Reiter, 2008);
2. for authors in improving and/or simplifying texts when, used to indicate whether they have appropriately targeted their intended audience.

Readability measures are supposed to enable anyone, without special knowledge or training, to determine the proportion of people who could comfortably read a piece of text. This assumes that those using the readability measures are aware of the ideal values they are attempting to obtain. Authors using these metrics may attempt to iteratively simplify technical and scientific documents to ensure they can be understood by wider audiences. However, many readability researchers advise against attempting to influence readability formulae in this way as modifying small amounts of text does not guarantee that texts are any easier to understand. They suggest that readability formulae should be used only for iterative feedback on the entire document (Klare,1984), though it should be possible to evaluate the impact on the readability score of specific changes to the text.

Although readability measures can be used as comparable objective measures of text difficulty, but they do not explicitly consider the content of the document. This means that they cannot, either, measure conceptual difficulty. By some formulae, Einstein's theory of relativity reads for ages 10-11. In addition, they cannot check the text is syntactically or semantically acceptable; "a man walks across the street" will score identically to "a street walks across the man". Indeed, both will score identically to "walks across a the man street". There is no semantic understanding embedded within the measures, the scores remain the same even if the text is scrambled. The readability formulae are also demonstrably inconsistent if used to consider changes to a specific piece of text. A particular modification to the text may improve readability according to some measures, and make it worse according to others. This is a consequence of the differences in weighting on the input factors. For example, the sentence, 'Further funding comprises an element of additional financing for those institutions which have a high historic unit cost' contains the word 'comprises'. If we substitute this word for the phrase 'is made up of', all the readability formulae will register more readable text, apart from the FOG index which will suggest this change made the sentence slightly harder.

2.1.1 Vocabulary

Oakland and Lane cite factors such as simplicity or familiarity as more effective means of measuring the difficulty of a word than by counting characters or syllables. They suggest that word difficulty can be determined by examining whether a word is challenging, unusual or technical, and suggest vocabulary as a text factor contributing to text difficulty. If the text is populated with difficult words then it becomes harder to read causing readers to complain about 'jargon'. Here, the word 'jargon' has a negative connotation implying unnecessary overuse of complex terms where the same scientific concepts could be expressed in non-technical terms without loss of understanding. The problem with unfamiliar words for a novice is that they may become part of the jargon they use as an expert and forget how to write for novice audiences. In scientific domains, terminology collections provide definitions for a large numbers of terms, each of which represent a specific concept. The meaning of these specialized semantic units may be difficult to deduce accurately for a novice reader. In some cases, a terminology collection may be provided alongside some documents to help in understanding the specialized documentation. This is particularly true for international standards (ISO) where the construction of the terminology should be a vital consideration for avoiding ambiguity and accurate application. A terminology collection may be essential both for novice readers hoping to get to grips with a domain and for those applying standards in order to do so accurately.

The European Association of Aerospace Industries (formerly AECMA, now ASD) developed Simplified Technical English (formerly AECMA Simplified English), a specification for aircraft maintenance, to ensure non-native speakers of English did not create potentially dangerous situations through misinterpreting documentation. The specification includes a dictionary providing a limited vocabulary for use in their documents. Each word in the dictionary has only one meaning, for example the word ‘drive’ is always used in the mechanical sense such as ‘the drive was faulty’. The word can never be used to describe a journey such as ‘the drive was boring’. Having a predefined description for a word can help avoid confusion. The terminological nature of such documents means that the terms are used disproportionately frequently throughout the documents, and more often than one would expect to encounter in everyday language. While, this means that anyone unfamiliar with the terminology would find the vocabulary used in the document hard to understand, it also means that terminology can be identified as words used frequently in the document but are unfamiliar in general use. From this perspective, vocabulary could be measured as the frequency of the words within the document in accordance to their familiarity in general language.

2.1.2 Syntax

The vocabulary, however, does not tend to exist in isolation, and to understand why text can easily become difficult we may also consider grammar. The vocabulary may be well-defined, yet included in overly verbose sentences. Furthermore, the relationships between the terms may hinder understanding. Oakland and Lane suggest that long and complex sentences, and unfamiliar syntactic structures, can confuse the reader. They include syntax as a text factor contributing to text difficulty.

Text is usually confusing due to ambiguities arising when information is not presented appropriately. Nearly all sentences are multiply ambiguous if taken out of context. Analyzing syntactic features, such as sentence length and appropriate punctuation does not always remove ambiguity. The Plain English Campaign offers rules and techniques that are intended to reduce ambiguity, improve understanding, increase reading speed and ease translation. In particular, the campaign provides an A-to-Z guide of over 1300 plain English substitutions for supposedly difficult words and phrases. According to this list, and depending on sentence structure, “essential” could be replaced with either “important” or “necessary” and “according to our records” could be substituted for “our records show”. A similar list of substitutions was also produced by ASD Simplified Technical English.

However, difficult words and phrases cannot always be so easily substituted. Words tend not to combine randomly or freely, rather they are used with preferred ‘friends’. According to Firth (1957) “you shall know a word by the company it keeps!” and this may be evident in combinations as certain kinds of ‘collocations’. Collocations demonstrate the preference for friends, and importantly have “distant” friends, i.e. there is significance to the distance and order between the collocating words: for example, the collocation between “bread” and “butter” is rather more frequently encountered as “bread and butter” than as “butter and bread”. In addition, expected synonyms may be largely excluded as friends, so while we have ‘strong tea’ we appear to have rather less by way of ‘powerful tea’. However, a reader unfamiliar with such constructions might not understand the precise meanings or variations. Put another way, individual words may not be particularly difficult but their combination can produce different meanings to those that component words might suggest. This is readily demonstrable in general language, where a ‘tyre’ is thought of as an object of a circular nature but ‘flat tyre’ does not refer to a flat circle, rather it indicates a lack of air pressure.

In specialised writing, collocations can become complex, providing a more specified meaning, but as a result become harder to understand. Consider, for example, ‘glass crack growth rate’. On a word-by-word basis this, should be easy to understand. However there are several possible interpretations, due to bracketing

(Pustejovsky et al., 1993) that might lead us even to consider a 'crack growth rate' made of 'glass'. ASD Simplified Technical English recommends writers avoid lengthy collocations by breaking them up. For example, instead of 'runway light connection resistance calibration', you should write 'calibration of the resistance on a runway light connection'. This unpacking of semantics helps to remove the ambiguity arising from bracketing and considerations of adjacency and dependency (Lauer, 1995) can help determine the roots of the semantic packing.

However, long collocations are only difficult to people who have not come across them before. Experts in a particular subject will find disambiguating their long terminological expressions considerably easier. Some linguists consider collocations to be 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. Smadja (1993) elaborated on the criteria for a collocation, describing them as recurring and cohesive domain-dependent lexical structures. Smadja used examples of 'stock market' and 'interest rate', and suggested how components can imply collocations, for example 'United' produces an expectation of 'Kingdom', 'Nations', or 'States'. It is these frequently combined linguistic expressions that develop into a processing unit, where many of the linguistic elements are ignored. The whole chunk is compressed and treated as one semantic unit. These units often develop into terminology with multiword units representing singular concepts. This is why experts familiar with the terms in a subject field find it easier to understand them.

When we read, we use semantic units to build a collection of concepts described within the text. It is with these blocks of semantic units that we form our interpretation of the text. An analysis of readability should consider collocations in the text and not just the individual words. Most researchers agree that collocations can be detected using statistical measures of association as they are sequences of words that co-occur more often than by chance, assuming randomness. It could be possible to consider syntax, in this case syntactic complexity, as measurable by considering the extent of collocations and the complexity of phrasing in the text.

2.1.3 Cognitive Load

When a significant amount of information is conveyed in a relatively small amount of text, the problem of cognitive load occurs. 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 amount of ideas expressed in the text contributes to cognitive load by increasing the work demanded of the reader by authors to interpret their text correctly. Perhaps cognitive load is measurable by examining the quantity of defined and undefined terms within short distances of each other. This will determine the workload required by a reader to process or interpret the text correctly.

2.1.4 Idea Density

So-called idea density manifests 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. While this poses no problem for specialists, it can often be intimidating for novices. Writers often expect the reader to make 'semantic leaps' (Halliday and Martin,1993) from existing understanding to understand particularly abstract ideas and conclusions, and this may lead to incorrect inferences. It is easy to confuse idea density with cognitive load, but where cognitive load is concerned with the number of ideas in the text, idea density refers to 'strength' or 'abstractness' of the ideas. Idea density is linked to vocabulary (terminology) in that an expert will find it easier to associate the content of the text.

Idea density may be related to considerations such as lexical cohesion (Hoey 1991), where repetition of a lexeme and its synonyms provides a structure for the reader to connect with. Repetitious patterns help readers form an understanding throughout the text. Sentence links and bonds enable summarization and allow consideration of the overall characterization of the text. If a large number of new, seemingly unrelated ideas are being introduced, this should be evident in low cohesion. Perhaps lexical cohesion can provide us with a measure of idea density.

2.2 Reader Factors

Text factors presume that difficulty is an artefact of text. However, different readers will have different views of the same piece of text. Reader characteristics may amplify or negate problems with difficult text. For a variety of reader factors identified by Oakland and Lane, we consider that it would be necessary somehow to capture and analyse the user's experience with prior documents as a proxy for reader knowledge, and that the capture and analysis would lead towards a personalized assessment for the document.

2.2.1 Background Knowledge

Although many readability metrics determine a grade level of an audience capable of understanding the text, they make no distinctions according to the background knowledge of the reader. Consider a general reader confronted in text discussing a 'muon'. The term is short and would be rated as simple by the current readability formulae. However most people would be unfamiliar with this term and only particle physicists are likely to know the term, its definition and related items. Oakland and Lane suggest that background knowledge contributes to text difficulty: a reader well-versed in a particular subject field should find the words rather more familiar. A longer word may only be difficult for a particular reader and certain shorter words may be more difficult to understand for wider audiences. The reader's familiarity with a word gives a much better indication of word difficulty than word length. Similarly, the difficulty of collocations is dependent on their level of familiarity with the reader. Long terminological phrases are only difficult to readers outside the domain.

Klare et. al (1955) described a series of studies conducted by the U.S. military showing how prior knowledge affected readability. In a manner similar to the Plain English Campaign, they simplified and changed the style of technical documents while experts ensured that all the technical terms were kept and that the intended message was not changed. The simplified versions resulted in faster reading speeds and greater retention of information, but differences were only noted in readers who were naïve in the subject. There was little observed benefit for the experts. Entin and Klare (1985) follow up these experiments to demonstrate that more readable text is beneficial for those with less knowledge and interest. Knowledge of a subject effectively 'drowns out' problems of difficult text. However, it is not easy to measure the amount of background knowledge required to differentiate between levels of knowledge; DuBay (2004) queries the measurement used in the conclusions from results of reading tests – is this a reflection on comprehension, prior knowledge, memory, or just the difficulty of the question used in the reading test? More generally, the reader's background knowledge needs to be captured and measured somehow.

Perhaps a terminology collection and its definitions can in some way, reflect the background knowledge required by a reader to interpret the text correctly. Knowing the precise meaning of certain words can help distinguish some of the ambiguity of the surrounding words. One way to measure background knowledge would be through the extent of use of terminology in the text with consideration of previous documents within the reader's experience.

2.2.2 Motivation and Engagement

Entin and Klare (ibid.) showed that more readable text is beneficial for those with less knowledge and interest. In another part of 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. Oakland and Lane characterize this as motivation and engagement. A study by Klare (1976) showed that experiments using readability formulae to simplify texts can be skewed by the interests and motivations of the reader: readability is more important when interest is low. Some researchers argue that the link between text comprehension and motivation is due to the extent of reading performed by the reader. Cox and Guthrie (2001) provide evidence that reading motivation predicts quantity of material read, and this in turn predicts text comprehension. In some ways, then, a readability system would need to ascertain whether a reader had demonstrated an interest in similar or related previous material. One way to measure motivation would be to examine a reader's history for similar documents, building on the measurement for background knowledge.

2.2.3 Language

Oakland and Lane identify language as another reader factor contributing to text difficulty. The process by which readers develop sufficient knowledge of a language is referred to as language acquisition and concerns familiarity with words and the development of the language capability. Researchers have shown that frequency is one of the strongest determiners in acquiring language, but have yet to explain how humans acquire the more abstract forms of linguistic knowledge. Bod et al. (2003) and Bybee and Hooper (2001) showed that frequency has an impact on comprehension and the development of language categories, and although it is widely assumed that grammar cannot be learned from experience alone, researchers working on collocations and distributional lexical semantics may produce interesting future insights. Ellis (1994) stated that frequency is a necessary component of theories of language acquisition but is not a sufficient explanation – otherwise, we would never get beyond the definite article in our speech. Distributional cues are useful for categorising high frequency items encountered in the identical contexts, and considerations of distributional lexical semantics pay strong heed to this, but these cues are less useful when considering lower frequency words.

We discussed in syntax how collocations become semantic units representing singular concepts. These collocations sometimes become quite ambiguous multi-word expressions. Research has shown that frequency is indispensable for dealing with these ambiguities. Ford, Bresan and Kaplan (1982) showed how subjects used innate statistical information to determine how sentences should be interpreted. 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. Diessel concluded that several psychological mechanisms such as information processing and analogy interact with frequency based mechanisms to develop linguistic structure. A person's grammar is an emergent linguistic structure developed from their use of language.

Familiarity is therefore fundamentally important for readability with the frequency of words and collocations relating to the difficulty a reader will have with them. The familiarity with words in language acquisition relates to the factors of vocabulary and background knowledge: while it is possible to ascribe an overall score for a word, perhaps as a measure of its rarity in discourse, words will have different familiarity for different readers. A difficult word for a novice is not always the same as a difficult word for an expert. However, beyond this, it is difficult to make a clear distinction between the reader factors of background knowledge and language unless we make consideration for the non-terminological elements of the text – measuring the reader's familiarity with words and collocations in general use.

2.2.4 Reading Fluency

Oakland and Lane identify reading fluency as the final reader factor contributing to text difficulty. As discussed in relation to language acquisition, the more text a person reads, the stronger their experience-based grammar becomes. This in turn results in a more fluent reader. Research has shown the importance of reading fluency in developing reading proficiency and differences in reading fluency can distinguish between good and poor readers. Stanovich (1991) showed how a lack of reading proficiency is a reliable predictor of reading comprehension problems. There is a strong correlation between reading fluency and reading comprehension with each aspect of reading fluency having a connection to text comprehension. Ehri and McCormick (1998) showed that factors such as knowledge of a large bank of high frequency words are needed for accurate word reading. Words are only analysed when they cannot be read from memory as sight words. This relates back to the reader factor of language: the reader's lack of knowledge of words will affect their reading fluency in that 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. Language acquisition experiments have shown that the categorization of word classes can be improved by incorporating phrasal boundaries. The correct placing of pauses around phrase boundaries contributes significantly to their meaning. For example, Rasinski (2003) used the following example to show how an ambiguity introduced into a string of words can produce interpretations that are either meaningful or nonsensical.

"The young man the jungle gym."

The majority of readers pause at 'man', rendering the phrase meaningless. However, if the reader pauses at 'young', they can construct the meaning and interpret the sentence. Schreiber (1980) suggested that fluent readers use morphemic, syntactic, semantic and pragmatic cues present in the text to organize it into meaningful phrases. This work relates to collocations and the text factor of syntax with frequent collocations used to decipher text. In the example, without any punctuation, the frequent collocation 'young man' is used to try and construct meaning from the phrase. In this instance, the collocation leads us to an invalid interpretation rendering the phrase meaningless. It is only with a phrasal boundary dividing the collocation that we can begin to interpret the phrase as it was intended.

Reading fluency is adversely affected by difficult text, Young and Bowers (1995) showed how the accuracy, speed and expressiveness of poor readers are more affected by text difficulty than average readers. Poor readers find difficult text harder to understand. Perhaps reading fluency can be addressed through the reader's familiarity with general language and the consideration of collocations and phrasal boundaries in a manner similar to the reader factor of language.

2.3 Commentary on Oakland and Lane

Oakland and Lane's framework has potential benefit in providing a more thorough analysis of readability than just counting words and sentences. Numerous research has shown familiarity is a good indicator of difficulty for both words and collocations. By using a frequency analysis we can provide an approximation for the difficulty of these expressions. We propose that an statistical analysis can identify problematic words and phrases in a document for a reader. However, any assessment is dependent on the reader and their experience with these expressions. If we assume the reader is an expert in a domain we can ignore words or phrases which would seem difficult to a novice. A terminology collection specific for the subject field can be used to find the known terminology in the text and therefore discount these words and phrases as troublesome.

We consider that the terminological component is key to measuring readability with unfamiliar words addressing **vocabulary** with the number of terms which the reader has previously encountered relevant for the **background knowledge**. Both **idea density** and **cognitive load** appear to relate to the introduction and packing of terms within the text. Furthermore, while **syntax** deals with the structuring amongst these, sometimes leading to collocations, **language** and **reading fluency** address the reader's familiarity with the syntax. Finally, **motivation and engagement** seems to be consistent with the frequency of previous encounters with the background knowledge.

3. Document Content Management System

We decided to develop a framework for automatically analysing readability by modelling Oakland and Lanes factors that contribute to text difficulty. GATE was selected as a basis and front-end for the implementation as it is established within the NLP community. The GATE interface allows for different "processing resources" to be executed in sequences in what is referred to as a pipeline; the user can order the running of these processing resources. A set of reusable processing resources for common NLP tasks is provided with GATE, packaged together to form A Nearly-New Information Extraction (ANNIE) system. We expanded the function of GATE into a system capable of document content management. Existing GATE plug-ins from ANNIE were used for the preliminary NLP tasks, leading into our newly devised processing resources. These additional resources and the results of analysis emerging from them will be described as follows:

- Terminology Lookup (3.1)
- Linguistic Term Finder (3.2)
- Keyword Extractor (3.3)
- Statistical Term Finder (3.4)
- SimpleText Analyser (3.5)
- Annotation Controller (3.6)

- Readability Analyser (3.7)
- Replacer (3.8).

The pipeline for these resources is shown in Figure 2, below, with brief descriptions of each component following to provide an indication of the approach. It should be noted that the readability analyser can be run at two separate points in the pipeline, the latter prior to committing changes.

Figure 2. Pipeline for the prototype document content management system.

3.1 Terminology Lookup

The Terminology Lookup plug-in analyses documents and annotates term entries. It uses an ISO 16642 compatible XML-based terminological mark-up file containing a snapshot of the terminology collection. The terminology is available in both English and French, potentially providing some assistance for translators also. Providing such a file containing terminology for different domains/applications is a possibility. In this instance, extant terms are annotated using our own collection, though in future this could interoperate directly with the iTerm TMS that has recently been populated with ISO TC37 terminology.

3.2 Linguistic Term Finder

The Linguistic Term Finder identifies candidate terms according to specified patterns of part of speech annotations (e.g. Jacquemin 2001:27) using the ANNIE POS tagger within GATE.

3.3 Keyword Extractor

The Keyword Extractor calculates distributions of frequency and weirdness as outlined by Gillam (2004). We use frequency information from the 100 million word tokens of the British National Corpus (BNC) to act as a reference corpus. The extent to which annotations are applied can be adjusted by modifying parameters for the distributions and their combinations.

3.4 Statistical Term Finder

The Statistical Term Finder takes input from the Keyword Extractor (3.3). This plug-in examines collocations of the keywords and identifies patterns occurring with statistical significance, following on from the work of Smadja (1993) and Gillam (2004). We use defined thresholds for identification: if a word consistently appears in the user-defined neighbourhood size above the threshold value, it is considered a potential new term. This can be undertaken iteratively (re-collocation). We are exploring automatic determination of this threshold in related work.

3.5 SimpleText Analyser

The SimpleText Analyser uses a thesaurus containing words and phrases identified as verbose, and hence deprecate, by either the Plain English Campaign or ASD Simplified Technical English . The thesaurus contains 1302 such entries, offering one or more preferred alternatives for each. The SimpleText Analyser identifies these phrases within the text and offers potential replacements for the expression. We are examining improvements to automation of this aspect.

3.6 Annotation Controller

The Annotation Controller is used to reduce the quantity of overlapping annotations being produced by prioritising some annotations over others. This component was added after discovering that conflicting suggestions for improvements were being produced by the components operating in parallel.

3.7 Readability Analyser

The Readability Analyser computes the number of words, syllables, sentences, characters and polysyllabic words contained within a document as required by current readability formulae. These values are used for the calculation of readability formulas such as the Kincaid formula, Flesch Index, SMOG, ARI and Fog Index.

3.8 Replacer

The Replacer substitutes the text in a document with user-selected SimpleText replacements (3.5). If no "best replacement" is selected, the text is left unchanged. Once the Replacer has finished, the whole procedure can be repeated. When the Readability Analyser is subsequently run, the effect the replacements had on the readability scores of the document are displayed and the user can decide whether any further replacements or additions to the subject terminology are appropriate.

3.9 Experiment and results

To demonstrate results of this analysis, two standards being developed within the LIRICS project, at various stages of the ISO process, have been analysed using this prototype system. The documents ‘Lexical markup framework (LMF)’ (at Draft International Standard stage) and ‘Syntactic Annotation Framework (SynAF)’ (at Working Draft stage) were chosen to show the output obtained from the various stages of the analysis. For the known terminology lookup we found that all known terms were annotated including those occurring with another term within the terminology annotation. For example, the known term ‘object language’ had another known term ‘object’ annotated within it. Definition were provided for both of the terms.

Multiple annotations were also found using the plug-ins for identifying terms (Keywords, Statistical and Linguistic). For example, the potential term ‘syntactic annotation’ also has a potential term, ‘annotation’, within it. Discovered terms often had known terms annotated within them. For example, the discovered term ‘dependency information’ has the existing term ‘information’ annotated within it. This new proposed term could then become an extension of the existing terminology. Decisions over the use of such relationships need to be considered. The numbers of known and discovered terms (total count) found in the two documents are detailed in Table 2.

Document	Known Terms	Discovered Terms
Lexical markup framework (LMF)	466	3712
Syntactic Annotation Framework (SynAF)	96	1125

Table 2. Number of known and potential terms in the ISO standards currently contained in the terminology database.

The ‘LMF’ document was roughly three times the size of ‘SynAF’, but appears to have substantially more terminological content. The discovered terms were investigated further to evaluate which could be considered as potential new terms. The terms highlighted by both methods were prioritised for consideration. Terms such as ‘syntactic annotation’, ‘annotation’, ‘SynAF’ and ‘morph’ were identified as items that may need to be defined. Further filtering of this list is required, but frequency information can be helpful here also; variations by part of speech can lead to duplications, for example for ‘SynAF’. Examples of discovered terms from SynAF are shown in Table 3. Additionally, the linguistic and statistical methods for discovering terms found numerous valid two word expressions that were regularly used. Examples of these are shown in Table 4.

Term	Linguistically Valid	Statistically Valid	Count
* annotation	N	Y	42
head	Y	N	33
value name	Y	N	22
partec	Y	Y	21
* synaf	Y	Y	19
value	Y	N	18
edge label	Y	N	14
syntactic annotation	Y	Y	13
mod	Y	N	11
morph	Y	Y	11
* synaf	N	Y	11
word	Y	N	11
* annotation	Y	Y	10
constituency	Y	N	10

Table 3. Examples of highly frequent discovered terms in 'SynAF', including duplications due to different parts of speech (*).

Term	Linguistically Valid	Statistically Valid	Count
sense class	Y	N	14
lexicon instance	Y	N	8
core package	Y	N	6
sense instance	Y	N	6
external system	Y	N	5
lemma class	Y	N	4
narrative description	Y	N	4
word forms	Y	N	4
affix class	Y	N	3
affix slot	Y	N	3

Table 4. Examples of frequent bigrams in 'LMF'.

There were also notable keywords (single words) identified as either linguistically or statistically valid and frequently used throughout the document. Examples of these are shown in Table 5. Some of these may easily be filtered out.

Term	Linguistic Discovery	Statistical Discovery	Count
LMF	N	Y	34
ISO	Y	N	27
subcategorization	N	Y	24
multilingual	N	Y	18
verb	Y	N	11
inflectional	N	Y	9
agglutination	Y	N	8
UML	N	Y	8

Table 5. Examples of discovered single-word terms in 'LMF'.

Discovered terms of increased length at lower frequencies indicate the existence of potentially highly complex expressions. The combination of the two methods of identifying potential new terms allows for readability issues caused by ambiguous bracketing to be highlighted. Such a readability issue can be demonstrated by the first item in Table 6, the "complex knowledge organization system":

1. [complex knowledge] [organization system]: an organization system for complex knowledge, simple knowledge is excluded?
2. [complex] [knowledge organization system]: a knowledge organization system that is somehow complicated?
3. [complex knowledge organization] [system]: the system is for an intricately arranged "knowledge organization"?

Table 6 also demonstrates term inclusion: "data category" is a term from "ISO 1087-2:2000 Terminology work - Vocabulary - Part 2: Computer applications" and "data category selection" is defined in "ISO 12620:1999 Computer applications in terminology - Data categories". We find 2 instances of "lmf data category selection procedures", which appears to extend this notion somehow. Correct interpretation, however, remains an exercise for the document author. The discovery of the "multi-layered annotation" and its "strategy", or perhaps the "annotation strategy" and its multiple layers, may also suggest the need for the correct interpretation to be made clear.

Term	**Linguistic Discovery**	**Statistical Discovery**	**Count**
complex knowledge organization system	Y	N	4
lmf data category selection procedures	Y	N	2
semantic predicate class section	Y	N	2
dual use mrd metamodel	Y	N	2
dual use mrd package	Y	N	2
multi-layered annotation	N	Y	3
multi-layered annotation strategy	Y	N	2

Table 6. Examples of potential multiword terms that were discovered in 'LMF'.

In analysis using a subset of the Plain English substitutions against a further document, ISO/DIS 12620, a report of substitutions for words and phrases deemed unnecessarily complex was produced. The first 200 replacements were analysed manually and it was found that 33 replacements were suitable. Every further instance of these substitutions was analysed throughout the rest of the document, 183 instances in total, to see if the replacements were appropriate in every instance. We found 65 potential replacements were valid. The suggestions and replacements are detailed in Table 7.

In the SimpleText analysis, it was found that some replacements were appropriate in every further instance such as "comprises", "in order to", "permissible" and "thus". However, some of the other words rarely had correct replacements and in particular 'application' and 'component' were never suitable again. In fact, the majority of the proposed SimpleText replacements were found not to be suitable. This left much room for investigation in how to focus the substitutions more accurately. To investigate the extent that this limited number of substitutions could influence the readability scores of the document the Replacer (3.8) plug-in was run. All the readability scores were slightly reduced except for the FOG and SMOG results which increased a little. It should be noted that Flesch also increased but this is because with this measure higher scores indicate more readable text. The increase in the FOG and SMOG scores can be attributed to the fact that some SimpleText replacements do not increase readability scores. In fact the number of words in a document can actually increase due to some of the replacements. The most common example of this occurrence is the substitution of "comprises" for "is made up of". Other replacements such as "important" for "essential" has no effect on readability scores whatsoever as the number of syllables and characters is identical. The readability scores before and after the replacements are shown in Table 8.

Phrase	Replacement	Occurs In Text	Replaced	% Correct
application	use	17	1	5.88%
by means of	by	2	2	100.00%
component	part	68	1	1.47%
comprises	is made up of	4	4	100.00%
consequence	result	1	1	100.00%
essential	important	2	2	100.00%
in conjunction with	with	2	2	100.00%
in order to	to	4	4	100.00%
various	different	10	4	40.00%
within	in	20	14	70.00%

Table 7. Replacements filtered from initial suggestions, with the number of times the replacements were correct throughout the rest of the document.

Score	Before	After
Kincaid	14.753	14.747
Flesch	28.534	28.611
FOG	*17.234*	*17.254*
SMOG	*15.432*	*15.447*
ARI	14.408	14.398

Table 8. Readability scores before and after the SimpleText process.

4. Weirdness Formula

Encouraged by the success of the using 'weirdness' for terminology extraction, we extended the notion to measure word difficulty. Currently, the majority of readability formulas use word length (by syllables or characters) as an indicator for word difficulty, we propose based on theories of language acquisition that word difficulty is dependent on familiarity. The difficulty of a word is dependent on how many times it has been previously encountered by a reader. Based on this premise, we used the frequency information form the BNC to determine the difficulty of a word. Here, the most frequent word is 'the' and therefore this is deemed the easiest word in English using our measure. Covington (2008) devised a measure for idea density which varied from Oakland and Lane's definition of the term. Covington gathered a collection of benchmark documents using the Google query "predicts U.S. inflation rate" and included four speeches by Federal Reserve chairmen. We compared our readability measure against Covington's and other common readability formula using

Covington's collection of documents. The performance of the measure in compared to the other measures is shown in Table 9.

Document	**Genre**	**Weirdness**	**Kincaid**	**FOG**	**ARI**	**SMOG**	**Covington**
Bloomberg, U.S. Leading Indicators	Popular	13.72	10.83	13.05	10.40	12.32	0.43
Wikipedia, Monetary Policy	Introductory	13.80	14.64	17.02	14.32	15.22	0.47
Associated Press, Fed Revises	Popular	12.88	9.82	11.60	9.10	11.30	0.47
USA Today, Greenspan predicts	Popular	13.09	10.68	12.59	11.18	10.91	0.48
Kitov, Exact Prediction	Scholarly	13.89	13.00	15.44	12.53	14.15	0.48
Wikipedia, Inflation	Introductory	13.93	13.88	16.24	14.00	14.59	0.48
Hyclak and Ohn, Wage Inflation	Scholarly	14.21	15.64	18.46	16.27	16.10	0.48
Investopedia, Trying To Predict Interest Rates	Introductory	14.03	12.61	15.06	13.23	13.50	0.49
Greenspan, to congressional committee 2005	Technical	13.99	14.02	16.04	14.02	14.41	0.50
Wright, Forecasting U.S. Inflation	Scholarly	13.73	13.01	16.72	13.29	15.07	0.50
Bernanke, speech 2008	Technical	14.59	16.26	17.98	17.62	15.55	0.50
Bernanke, report to congress	Technical	14.55	16.66	18.94	17.70	16.31	0.51
Stockman, Dollar Depreciation	Technical	13.64	14.42	16.59	14.99	14.86	0.51
Greenspan, speech 2005	Technical	14.07	15.17	17.30	15.43	15.35	0.52
Correlation			**0.87**	**0.86**	**0.87**	**0.84**	**0.40**

Table 9. Number of known and potential terms in the ISO standards currently contained in the terminology database.

We found that the weirdness measure consistently correlated with the other readability measures confirming that word frequency is a good indicator of word difficulty. As noted by Covington, their idea density formula did not correlate with the other readability formula or our new measure. Due to the success of these results, we have implemented our weirdness measure for the Open Office suite of applications. The measure can be found in an extension for the Open Office word processor application called 'Writer'. The extension performs all the traditional readability metrics in addition to our new weirdness measure. Unlike the other metrics, which require a minimum amount of text, the weirdness measure can be performed on every sentence in the document. This means we can tell our users the easiest and most difficult words in the document. The extension is freely available from the open office website and can be downloaded from the following link: http://extensions.services.openoffice.org/project/ReadabilityReport

5. Conclusions

While typical readability formulae can be useful for comparative measurements, they largely lack consideration of external factors that may make a text more or less easy to understand. Only the two factors of syntax and vocabulary identified by Oakland and Lane as contributing to text difficulty are addressed by the readability formulas. No reader factors such as background knowledge or the reading ability of the reader are considered. The same text can have various levels of difficulty for different readers. Even the factors they do examine are not probably addressed with Oakland and Lane stating that for vocabulary, word familiarity should be used as a means of measuring word difficulty rather than word length. In addition, while readability formulae consider short sentences, little by way of other lexical or syntactic features tend to be examined. In our work, we primarily consider the terminological complexity and how this might be useful as a means to compute the readers understanding of a new text. Avenues for further research are likely to include: (1) improving SimpleText performance through consideration of local contexts (lexical and/or grammatical); (2) consideration of cognitive load through semantic distance; (3) enhanced statistical (term) detection by improving discrimination of collocation patterns, with related investigations currently being undertaken on the Enron email corpus; (4) production of readability metrics that incorporate the results of the approach outlined, taking account of additional background knowledge that can be provided alongside the document. For example, complexity based on the number of syllables fails to take into account existence of definitions. One outcome from our work, in due course, will be alternative measures for readability that takes account of such considerations.

Additional work needs to be done to address the remaining text factors of cognitive load and idea density. One possible method is to use lexical cohesion as described by Hoey (1991). Lack of cohesion increases the work required by the reader to interpret the text correctly and can be analysed using repetition of concepts as a guide to consistency in the document. WordNet and other thesauri can be used to find alternate labels for the same concepts. Further work will consider indicative summaries (Benbrahim and Ahmad, 1995) and Kintsch's (1998) theory of more cohesive concepts. A text should only have a few prominent ideas and concepts, if there are too many then the document is not cohesive. We aim to use principles of cohesion to devise a measure for the cohesiveness of a document.

Further work needs to be done to explore the two remaining reader factors, reading fluency and motivation that have not been addressed yet. Reading fluency is affected by phrasal boundaries and word familiarity with difficult text having a greater effect on those with poor reading fluency. We aim to evaluate the benefits of simple English replacements, word frequency in the British National Corpus and phrasal boundaries on those with poor reading fluency. A possible implementation of motivation involves examining the documents the reader has previously read to ascertain their interest in new material. The amount of previous material read in a subject field is an indicator of motivation. The results from this work could potentially inform an overall readability metric incorporating as many of these elements as the user chooses, based on the availability of information about the reader and the nature of the text.

These methods for readability do not apply to just human readers, software applications seeking to access the semantic content expressed in text also stand to benefit. For Natural Language Processing (NLP) systems, their abilities as machine "readers" of text are clearly relevant to how effectively they can process text, and indications of, for example, ambiguities within the text are beneficial to both human and machine readers alike. By improving the readability of text and incorporating factors which help human readers understand text, we hopefully increase the likelihood of effective automatic processing – improved machine-readability.

6. References

ASD Simplified Technical English. Available at http://www.simplifiedenglish.net (accessed: 30 September 2009)

Bod, R., Hay, J. and S. Jannedy (Eds.) (2003). *Probabilistic linguistics*. Cambridge, MA: MIT Press.

Benbrahim, M. and K. Ahmad (1995). "Text Summarisation: The role of lexical cohesion analysis". *The New Review of Document and Text Management*, 1, 321-335.

Bybee, J. and P. Hooper (Eds.) (2001). *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins.

Covington, M. A. (2009). "Idea density: a potentially useful characteristic of retrieved documents". *Proceedings, IEEE SoutheastCon 2009*.

Cox, K. E. and J. T. Guthrie (2001). "Motivational and cognitive contributions to student' amount of reading". *Contemporary Educational Psychology*, 26, 116-131.

Cunningham, H., D. Maynard, K. Bontcheva and V. Tablan (2002). "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications". In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics(ACL'02),* Philadelphia, July 2002.

Dale, E. and J. S. Chall (1949). "The concept of readability". *Elementary English,* 26, 23.

Diessel, H. (2007). "Frequency Effects in Language Acquisition, Language Use, and Diachronic Change". *New Ideas in Psychology*, 25, 2, 108-127.

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

Ehri, L. C. and S. McCormick (1998). "Phases of word learning: Implications for instruction with delayed and disabled readers". *Reading and Writing Quarterly: Overcoming Learning Difficulties*, 14, 2, 135-164.

Ellis, N. (1994). "Implicit and explicit language learning: An overview". In N. C. Ellis (Ed.) *Implicit and explicit learning of languages*, San Diego, CA: Academic Press, 1-32.

Entin, E. B. and 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.

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

Flesch, R. (1948). "A new readability yardstick". *Journal of Applied Psychology,* 32, 221-223.

Ford, M., J. Bresan and R. M. Kaplan (1982). "A competence-based theory of syntactic change". In J. Bresan (Ed.), *Mental representations and grammatical relations*, Cambridge, MA: MIT Press, 727-796.

Gillam, L. (2004). "Systems of concepts and their extraction from text". Unpublished PhD thesis, University of Surrey

Gunning, R. (1952). *The Technique of Clear Writing*. New York: McGraw-Hill.

Halliday, M. A. K. and J. R. Martin (1993). *Writing Science: Literacy and Discursive Power*. London : Falmer Press.

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

Jacquemin, C. (2001). *Spotting and Discovering Terms through Natural Language Processing.* The MIT Press.

Kincaid, J.P., R. P. Fishburne, R. L. Rogers and B. S. Chissom (1975). “Derivation of new readability formulas (automated readability index, Fog count and Flesch reading ease formula) for navy enlisted personnel”. *Research Branch Report 8-75*, Naval Air Station, Memphis, TN.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

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

Klare, G. R. (1963). *The measurement of readability*, Ames, Iowa: Iowa State University Press.

Klare, G. R. (1976). “A second look at the validity of the readability formulas”. *Journal of Reading Behaviour*, 8, 159-152.

Klare, G. R. (1984). Readability. In P. D Pearson, R. Barr, M. L. Kamil, and P. Mosenthal (eds.) *Handbook of Reading Research.* New York: Longman, 681-744.

Klare, G. R., Mabry, J. E. and Gustafson, L. M. (1955). “The relationship of style difficulty to immediate retention and to acceptability of technical material”. *Journal of educational psychology,* 46, 287-295.

Lauer, M. (1995). *Designing Statistical Language Learners: Experiments on Noun Compounds*. Ph.D. thesis, Macquarie University, Sydney, Australia

McLaughlin, H. (1969). “SMOG grading - a new readability formula”. *Journal of Reading*, 22, 639-646.

Oakland, T. and H. B.Lane (2004), “Language, Reading, and Readability Formulas: Implications for Developing and Adapting Tests”. *International Journal of Testing*, 4, 3, 239-252.

Pustejovsky, J., S. Bergler and P. Anick (1994). “Lexical Semantic Techniques for Corpus Analysis”. *Computational Linguistics,* 19, 2, 331-358.

Rasinski, T.V. (2003). *The fluent reader: Oral reading strategies for building word recognition, fluency, and comprehension*. New York: Scholastic.

Schreiber, P. A. (1980). “On the acquisition of reading fluency”. *Journal of Reading Behaviour*, 7, 177-186.

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

Smith, E. A. and R. J. Senter (1967). “Automated readability index”. *AMRL-TR,* 66-22. Wright-Patterson AFB, OH: Aerospace Medical Division.

Stanovich, K. E. (1991). “Word recognition: Changing perspectives”. In R. Barr, M. L. Kamill, P. Mosenthal and P. D. Pearson (Eds.), *Handbook of reading research* (2, 418-452). New York: Longman.

The Plain English Campaign. Available at http://www.plainenglish.co.uk (accessed: 30 September 2009)

Williams, S. and E. Reiter (2008). “Generating basic skills reports for low-skilled readers”. *Journal of Natural Language Engineering*, Cambridge University Press Preprints.

Young, A and P. G. Bowers (1995). “Individual difference and text difficulty determinants of reading fluency and expressiveness”. *Journal of Experimental Child Psychology*, 60, 428-454.