# From corpus-based collocation frequencies to readability measure

**Nikolaos K Anagnostou and George R S Weir**
**Department of Computer and Information Sciences**
**University of Strathclyde**
**Glasgow G1 1XH**

## 1. Introduction

This paper provides a broad overview of three separate but related areas of research. Firstly, corpus linguistics is a growing discipline that applies analytical results from large language corpora to a wide variety of problems in linguistics and related disciplines. Secondly, readability research, as the name suggests, seeks to understand what makes texts more or less comprehensible to readers, and aims to apply this understanding to issues such as text rating and matching of texts to readers. Thirdly, collocation is a language feature that occurs when particular words are used frequently together for other than purely grammatical reasons.

The intersection of these three aspects provides the basis for on-going research within the Department of Computer and Information Sciences at the University of Strathclyde and is the motivation for this overview. Specifically, we aim through analysis of collocation frequencies in major corpora, to afford valuable insight on the content of texts, which we believe will, in turn, provide a novel basis for estimating text readability.

## 2. Corpus Linguistics

Corpus linguistics can be defined as "…the study of language based on examples of 'real life' language use" (McEnery and Wilson, 2001). As we can see from the definition, corpus linguistics in itself is not a branch of linguistics, like syntax, semantics and so on. Rather, it is a methodology and technique for language study that can be used in any branch of linguistics.

Corpus linguistics is an approach that is enjoying a considerable renaissance since the 1980s, after a long period of unpopularity in the 1960s and the 1970s. Coupled with the great advances in computer power in the last two decades, corpus linguistics has opened up new opportunities to study language and its mechanisms, both empirically and objectively, in ways not available to early linguists. At the heart of this methodology lies the corpus.

According to McEnery and Wilson (2001), "any collection of more than one text can be called a corpus: the term corpus is simply the Latin for 'body', hence a corpus may be defined as any body of text" (p. 29). Although simple, this definition is not sufficiently comprehensive, as it misses additional meaning that the term 'corpus' carries in modern linguistics. For this reason, the authors provide four additional, corpora-specific characteristics (McEnery and Wilson, 2001):

1. **Sampling and representativeness:** One cannot possibly collect all the texts of a language. The text population for languages like English is huge and new utterances are created every day. For this reason, corpora are based on sampling. In addition, when studying the variety of a language, corpora need to "maximally representative" of that variety, in order to provide a picture as accurate as possible and avoid being skewed.
2. **Finite size:** Corpora tend to have a finite size e.g. 1,000,000 words. As soon as a corpus reaches its word goal, collection stops and the corpus does not increase in size any further. The only exceptions to this are the so-called monitor corpora, like COBUILD, which are open-ended entities and constantly increase in size as new samples are added.

3. **Machine-readable form:** Nowadays it is taking for granted that corpora are "machine readable", that is they exist in an electronic format. Such corpora have the following advantages: they can be searched and manipulated in ways that are not possible for corpora in other formats; they can be enriched with a lot of useful information or, in other words, be annotated.
4. **A standard reference:** There is an implicit understanding that a corpus functions as a standard reference for the variety of language it represents. For example, the Brown corpus is regarded as the reference for written American English. Corpora provide a common framework for various studies, so that results can be compared and data-driven differences between studies can be minimised.

### *2.1 Types of Corpora*

A corpus is always designed with a specific purpose in mind, which in turn characterises the corpus itself. What follows is a list of some of the commonly used corpus types (Hunston, 2002, p. 14):

- **Specialised corpus:** A corpus of texts of a specific type e.g. newspaper editorials, scientific articles and so on. They are used to investigate a particular type of language. A well-known example of a specialised corpus is the Michigan Corpus of Academic Spoken English (MICASE), which focuses on spoken English in U.S. academia.
- **General corpus:** A corpus of texts of many types. They can include written or spoken material, or both, and tend to include a range of texts as wide as possible. General corpora are usually much larger in size than specialised ones and are commonly used as reference sources for general language studies. A well-known example of such a type is the British National Corpus (BNC), consisting of 100,000,000 words.
- **Comparable or translation corpora:** Two or more corpora in different languages or different varieties of the same language. They are mainly used to identify similarities or differences between the languages or varieties compared. An example of such a type is the International Corpus of English (ICE) that holds 1,000,000 words from every variety of English it includes.
- **Parallel corpora:** Two or more corpora in different languages, each containing material that has been translated from one language into the other. An example is the Minority Language Engineering Project (MILLE) that contains parallel aligned Panjabi-English texts.
- **Learner corpus:** A corpus of this type consists of a collection of texts produced by learners of a language. The purpose of such a corpus is to identify differences between the learners themselves and between learners and native speakers of the language. Such a corpus is the International Corpus of Learner English (ICLE).
- **Monitor corpus:** A corpus designed to track changes in a language. As stated before, monitor corpora constantly increase in size as new texts are added. A well-known example of such a type is the COBUILD or Bank of English corpus.

### *2.2 Uses for Corpora*

Corpora have many diverse applications in language studies. Here we will present a selection of some of the most important of these applications, from the perspective of language studies and the perspective of language engineering

In the context of linguistic studies, corpora are used in (McEnery and Wilson, 2001):

- Speech research.
- Lexical studies.
- Grammar and syntax.
- Semantics.
- Language teaching.

In the context of language engineering, corpora have found applications in:

- Part-of-speech analysis.

- Automated lexicography.
- Parsing[1].
- Multilingual corpus exploitation (machine translation, cross-lingual information retrieval etc.).

### *2.3 The British National Corpus*

The BNC is a key ingredient for the development of our research. Since the purpose of this project is to use a language variable like collocation frequency as a predictor of semantic difficulty in order to create a readability formula for the English language, we needed a corpus that is:

1. In British English, for this is the target language, and monolingual, because we had to derive collocation frequency data from one and only language.
2. General and based on sampling, because the aim is to create a formula with a range as wide as possible and not applicable only to a specific genre or sublanguage[2], so maximum representativeness in the language data is essential.

As described in its reference guide (Burnard, 2000, p. 3), the BNC is:

- **Sample-based**: it consists of text samples no longer than 45,000 words in general.
- **Synchronic**: it includes imaginative texts from 1960, informative texts from 1975.
- **General**: not specifically restricted to any particular subject field, register or genre.
- **Monolingual**: the text samples included in the corpus are substantially the product of speakers of British English.
- **Mixed**: it contains both spoken and written language examples.

So, the BNC meets all the requirements outlined above and, for our purposes, is well suited as a source of collocation frequency data.

## 3. Readability

### *3.1 Introduction*

The field of readability research is very active. As Klare (1984, p. 682) states, well over a 1000 readability references can be found in the relevant literature. In addition, more than 200 readability formulas exist today.

In this section, we account for some of the core definitions of readability and outline the approach that we prefer. We then list the factors that researchers have found to be most influential in affecting readability. Then, we focus on readability formulas, with a brief description of the most popular ones in use today. We also discuss the main uses for readability formulas and summarize their flaws.

### *3.2 What do we mean by readability?*

There is little agreement regarding the exact definition of readability. Before venturing to the more formal definitions, we might suggest that readability is what makes one text more difficult or easier to understand than others. According to Klare (1963), readability is "the ease of understanding or comprehension due to style of writing". This definition focuses on writing style, in contrast to factors like format, features of organisation and content (cf. DuBay, 2004).

In contrast, McLaughlin takes into account the importance of specific reader characteristics, such as reading skill, motivation, relevant knowledge, and how these interact with the text.

---

[1] The procedure of identifying higher level syntactic relationships in a text, e.g. noun phrases, verb phrases etc., is known as parsing (McEnery and Wilson, 2001, p.53).

[2] A language used to communicate in a specialized technical domain or for a specialised purpose, for example, the language of weather reports, drug interaction reports, etc. Such a language is characterised by the high frequency of specialised terminology and often also by a restricted set of grammatical patterns (source: www.essex.ac.uk/linguistics/clmt/MTbook/HTML/node98.html).

Thus, he describes readability as "the degree to which a given class of people find certain material compelling and comprehensible" (McLaughlin, 1969, cited by DuBay, 2004).

The definition that seems to be the most comprehensive, is the following (Dale and Chall, 1948): "In the broader sense, readability is the sum total (including interactions) of all the elements with a given piece of printed material that affects the success which a group of readers have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting."

If we analyse this definition further, according to Klare (1963), it follows that the main functions of readability are:
1. To indicate legibility of the printed material as well as its layout or typography.
2. To indicate ease of reading due to the interest-value or the aesthetics of writing.
3. *To indicate ease of understanding and comprehension due to the style of writing.*[3]

As we can see, the main functions of readability map well to definitions given at the beginning of this section, except for the first one, regarding legibility. In the past, the boundaries between legibility and readability were much less clear than today and the two terms were used interchangeably (a quick web search shows that sometimes this is often still the case). However, it is important to bear in mind that they denote different things.

Legibility research is concerned with the visual presentation of information (Tekfi, 1987) and focuses mainly on typeface and format factors. In contrast to that, readability research studies linguistic factors like word and sentence length. Both share the same objective, to ascertain the degree of reading ease of a piece of text and eventually find ways to improve it, but their approaches are totally different. Having said all that, the definition given by Klare (1963), has become the most commonly accepted meaning for readability, and is the one we adopt henceforth.

### *3.2.1 Factors that influence readability*

In a seminal work on readability research, Gray and Leary (1935) identified more than 200 variables that affect readability, and grouped these into four categories:

1. Content (judged most significant)
2. Style (slightly less significant)
3. Format (third in significance)
4. Features of Organisation (least significant)

Their research showed that the most important of these categories were content and writing style, followed by format and "features of organisation"[4]. Figure 1 illustrates these four basic elements of reading ease.

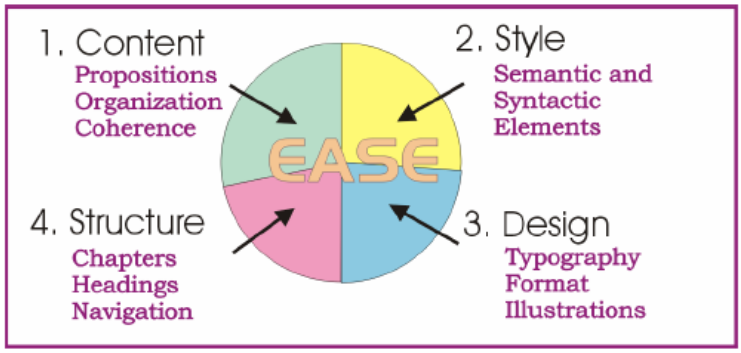


**Figure 1: The four basic elements of reading ease. (from DuBay, 2004)**

[3] The emphasis is ours.
[4] The term refers to the structure of a text in terms of chapters, sections, headings, etc.

As an aside, we might mention that readability was used as a general term in Gray and Leary's work, encompassing variables that are outside the current domain of readability research. Variables like typography and format today fall under the heading of legibility, not readability.

A significant finding was that of the four categories, only style - and variables related to it - could be measured statistically. The authors consequently characterised 64 style variables related to reading difficulty and used correlation coefficients to identify the best readability indicators. The factors with greatest impact were the following (DuBay, 2004, p.18):

1. Average sentence length in words.
2. Percentage of "easy" words.
3. Number of words not known to 90% of sixth-grade students.
4. Number of "easy" words.
5. Number of different "hard words".
6. Minimum syllabic sentence length.
7. Number of explicit sentences.
8. Number of first, second, and third-person pronouns.
9. Maximum syllabic sentence length.
10. Average sentence length in syllables.
11. Percentage of monosyllables.
12. Number of sentences per paragraph.
13. Percentage of different words not known to 90% of sixth-grade students.
14. Number of simple sentences.
15. Percentage of different words.
16. Percentage of polysyllables.
17. Number of prepositional phrases.

Gray and Leary's work stimulated a research explosion into finding the "perfect" formula and influenced most of the readability formulas in use today.

### *3.3 Readability formulas*

One common approach to predicting readability is the usage of readability formulas[5]. These are mathematical equations, constructed by linguists and readability researchers usually through regression analysis (McLaughlin, 1969), to help them gauge the difficulty and complexity of a given piece of text. The most frequently used formulas were created in the period from the 1930s to the 1970s and were constructed with a view to easy manual application. This is one reason why such formulas tend to contain very few variables. With the explosive growth of computers in the last three decades, most readability formulas are now computerised.

Readability formulas measure certain textual characteristics that are quantifiable. Such characteristics are usually described as "semantic" if they concern the words used and "syntactic" if they have to do with the length or structure of sentences. The two factors most commonly used in readability formulas are vocabulary difficulty, measured by either word difficulty or word length, and average sentence length, since a multitude of studies have proven them to be strongly associated with comprehension (Dave and Chall, 1995, p. 81). It is important to note that except for these surface-level features of texts, there are other variables that affect readability, like content and the reader's abilities, but these cannot be measured mathematically and for that reason are not included in readability formulas.

Normally, readability formulas return an estimate of a text's difficulty in terms of grade levels[6]. That is, the years of schooling needed to be able to comprehend the text. The grade-level scale was adopted because it provided a way to "compare reader's ability levels to the

[5] Examples of other ways of assessing readability are graphs, like the Fry Readability Graph, and text levelling, which uses qualified judges to determine the difficulty of texts (DuBay, 2004, p.35 & p.45).

[6] Grade level scores can be assigned to individuals as well, based on a typical reading test. In this case, such a score for an individual means that he or she reads as well as some normative group (Klare, 1984).

difficulty levels of written material" (Klare, 1984, p. 718) and thus guide teachers in the selection of appropriate material for their students. The outputs from such readability formulas tend to be more accurate in lower grade levels, because as the level increases, content, instead of writing style, becomes the deciding factor for readability (Klare, 1984, p. 730).

### 3.3.1 Developing a readability formula

How are readability formulas created? As expected, readability formulas are not the result of divine revelation but the product of a structured procedure that uses statistical techniques. Before describing the steps of the aforementioned procedure, it would be useful to give an explanation of two terms that are essential to the process and very common in readability research; *criterion passages* and *the cloze procedure*.

Criterion passages are sets of passages of increasing difficulty. Their difficulty is determined by independent checks of comprehension such as cloze, multiple choice tests or expert judgement[7] (also called text levelling). Most commonly, these passages are given a grade level. Criterion passages are used as a baseline against which a formula is standardised and cross-validated. A useful analogy is the creation of an equation in physics: The equation is derived from data regarding a natural phenomenon and afterwards its validity is checked by how accurate predictions it can provide for other aspects of the same phenomenon.

The cloze procedure is a deletion test, intended to measure a reader's understanding of a text. It uses a passage with words deleted at regular intervals (usually a five-word interval) and requires the subjects to fill in the blanks. The percentage of correctly entered words is the cloze score. The higher this percentage, the more readable the text is considered (Dale and Chall, 1995, p. 55). Cloze scores are converted to grade levels in the following manner: Assuming that a cloze score of *x* per cent signifies comprehension, the grade place of a passage is the average grade place of the subjects who achieved this score on cloze tests constructed from the passage.

With these two terms explained, it is time to see the steps needed to produce a readability formula. According to Klare, these steps are (Klare, 1984, p. 704):

1. Choose (or develop) a large set of criterion-text passages of varied content and measured difficulty (typically, comprehension) values.
2. Identify a number of objectively stated language variables.
3. Count the occurrences of the language variables in the criterion passages and correlate them with the difficulty values of the criterion passages.
4. Select the variables with the highest correlations as potential indexes of readability.
5. Combine the indexes, using multiple-regression techniques, into a formula.
6. [Optional] Cross-validate on a set of new criterion passages.

Our next section provides a brief overview of the some of most popular readability formulas in use today.

### 3.3.2 Brief overview of the main readability formulas

Out of the vast collection of more than 200 readability formulas, 5 are been singled out below, due to their popularity and their influence on subsequent work. The first three are milestones from the early period of readability studies. The last two are examples of more recent formulas[8].

.

**Flesch Reading Ease formula**

[7] It is important to note that that these tests attempt to *measure readability*, as objectively as possible, since they require human participation for their completion. This is in contrast with the function of readability formulas, which try to *predict readability* and do not require testing readers.

[8] Another interesting example of a modern readability measure is the Lexile framework, published in 1988 by Stenner et al. It uses average sentence length and average word frequency found in a corpus of 600 million words. It is not covered here, because the actual formula that the framework uses is not available.

This most popular of formulas was published by Rudolph Flesch in 1948, in an attempt to revise his first formula, in the face of various criticisms. It uses two variables, the number of syllables and the number of sentences for each 100-word sample (DuBay, 2004, p. 20). The formula returns a score on a scale of 1 to 100, with 30 being the most difficult and 70 being easy. A score of 100 denotes a passage that can be understood by people who are barely literate.

The Flesch Reading Ease score is given by the following equation:

$$RE = 206.835 - 1.015ASL - 84.6ASW$$

Where:
*RE* = reading ease (in a scale of 1 to 100).
*ASL* = average sentence length (the number of words divided by the number of sentences).
*ASW* = average number of syllables per word (the number of words divided by the number of sentences).

This formula was later revised by other readability researchers (Kincaid et al., 1975, cited by Klare, 1984, p. 692) in a study commissioned by the U.S. Navy, in order to provide grade level scores. The adapted formula is known as the Flesch-Kincaid or Flesch Grade-Scale formula.

**The Dave-Chall formula**
This is another very influential formula, published by Edgar Dale and Jeanne Chall in 1948. It was designed in such a way so as to correct certain flaws of the Flesch Reading formula. The formula uses two variables, average sentence length and a percentage of difficult words. That is, words not in the Dale-Chall list of 3000 easy words, 80 percent of which are known to fourth-grade readers (DuBay, 2004, p. 23).

The Dave-Chall formula is the following:

$$Score = 0.1579PDW + 0.496ASL + 3.6365$$

Where:
Score = reading grade of a reader who can answer 50 percent of the test questions on a passage.
PDW = Percentage of Difficult words (words not in the Dave-Chall word list).
ASL = as in the previous formula**.**

This formula has been recently revised and re-published (Dale and Chall, 1995). The new Dave-Chall formula has taken into account many of the readability research findings in the almost 50 years since the publication of the original and has greater predictive power than the original.

**Gunning Fog Index**
The Fog Index is another commonly used readability measure, published by Robert Gunning in 1952. It became popular due to its ease of use. The formula is based on two variables, average sentence length and the number of words with more than two syllables per 100 words (DuBay, 2004, p. 24). The output of the Fog Index is a grade level score.

The Fog Index is given by the following equation:

$$GL = 0.4(ASL + hard\ words)$$

Where:
*GL* = grade level.
*ASL* = as in the previous formula.
*hard words* = number of words with more than two syllables per 100 words.

**The SMOG Formula**

Published by G.H. McLaughlin in 1969, the SMOG formula is one of the simplest formulas to use and was based on the idea that semantic and syntactic difficulty *predictors should be multiplied*, instead of added. The formula uses only one variable, the number of polysyllable words (>2 syllables) in 30 sentences. This formula returns grade level scores.

The SMOG Formula is:

$$GL = 3 + \sqrt{PC}$$

Where:
*GL* = grade level
*PC* = polysyllable count (number of words with more than 2 syllables in a 30 sentence sample)

Due to the stricter criterion scores[9] used in the development of the SMOG formula, the grade levels it predicts are consistently at least two grades higher than the Dave-Chall formula (McLaughlin, 1969).

**The Bormuth Mean Cloze formula**
This readability measure is considered to be one of the most accurate. It was developed by John Bormuth in 1969 and was later adapted by the U.S. College Entrance Examination Board in 1981, for use in the Degrees of Reading Power (DRP) system (Klare, 1984). It uses three variables, average sentence length in words, average word length in characters and the number of words on the original Dale-Chall list of 3000 words. The formula in its original form returns a mean cloze score.

The original Bormuth Mean Cloze (BMC) formula is:

$$\begin{aligned} Readability(R) = {} & 0.886593 - 0.083640(LET/W) + 0.161911(DLL/W)^3 \\ & - 0.021401(W/SEN) + 0.00577(W/SEN)^2 - 0.000005(W/SEN)^3 \end{aligned}$$

Where:
*R* = mean cloze score
*LET* = letters in passage *X*
*W* = words in passage *X*
*DLL* = Dale-Chall list words in passage *X*
*SEN* = sentences in passage *X*

The mean cloze score that the formula returns is transformed into DRP units by using of the conversion formula below:

$$DRP = (1 - R) \times 100$$

The units range from 30 (very easy) to 100 (very hard).

It is important to note two things about the BMC formula that distinguish it from other readability measures:

1. While most of the readability formulas use two variables (mainly word and sentence length), the BMC formula uses three (two for semantic and one for syntactic difficulty).
2. *It introduces curvilinearity*. Bormuth's studies confirmed that the correlation of the formula variables with text difficulty changes in upper grade-levels, producing a curve when plotted on a graph (DuBay, 2004, p. 43).

[9] The scores of the tests used for measuring comprehension were stricter than the ones used for the other formulas.

### 3.3.3 Applications

Readability formulas have a wide range of applications. They can be generally grouped into two categories:

1. Selecting material appropriate for the targeted audience's reading ability.
2. Simplifying texts.

The main usages of readability formulas fall under the first category. They are frequently used in education by teachers, in order to find the material that is most appropriate for their students in terms of difficulty[10]. Readability formulas are also used to check publications in sectors like legislation, health information, technical manuals etc., in order to ensure that the published material is as easy to understand as possible, as failure to understand such documents can make the difference between life and death in many situations[11]. Finally, readability formulas have been used to evaluate the readability of scientific journals (Tekfi, 1987, p. 271). Information scientists have found that the difficulty of scientific journals is constantly increasing and readability formulas can be used to check this trend.

Readability formulas are have also been used by writers, editors and publishers in order to adapt a text to a given level of difficulty, through a process of writing, rewriting and revising (Chall and Dale, 1995, p. 49). That is, readability formulas can function as rules for writing or rewriting material to match a certain level of reading difficulty. It has to be noted though that many readability researchers have advised *against* the use of formulas in such a way, warning that the lowering of readability scores does not always produce texts that are easier to comprehend. Instead, the suggestion is that readability formulas should be kept out of the writing process itself and used only for feedback, according to the following cycle (Klare, 1984, p. 730): Write → Apply formula → Revise → Apply formula…

### 3.4 Criticisms

Even though these formulas are useful and objective measures of text difficulty, it is important to bear in mind that they have flaws. The field of readability research has witnessed several, quite heated debates about the shortcomings of readability formulas. Some of the main limitations of readability formulas that have been identified in the relevant literature are the following:

- **They cannot measure conceptual difficulty**: No formula takes into account the content of the document being evaluated. For example, by some formula's calculations, Einstein's theory of relativity reads at a 5th grade level (U.S. State Department, 1998).
- **They cannot check incomprehensibility of expression**: Readability scores remain the same even if the text is scrambled. For example, the phrase "Mary had a little lamb" will have the exact same score with the phrase "had lamb little Mary a". It is clear that the second phrase is incomprehensible but readability formulas are unable to detect that.
- **There is discrepancy in the results of readability formulas for the same text**: For example, the scores for the closing paragraph of this subsection are (in grade levels):
  Flesch Kincaid: 16.7
  Gunning Fog Index: 22.6
  SMOG Grading: 18.5
  These differences are caused by the different variables and different criterion scores used by different formulas, but are can still be perplexing.
- **They assume that all readers are alike**: Readability formulas make no distinctions based on reader's characteristics. That is, they take no account of differing purposes, maturity and ability of readers (Redish, 2000, p. 134).

[10] This has been labeled the problem of optimal difficulty, for the level of reading difficulty desired changes if we are talking about independent instead of assisted reading.

[11] Dubay (2004) states that 79 to 94 percent of child safety seats are installed improperly, due to poor comprehension of the installation manuals, a fact that has aggravated the rate of traffic accident related infant deaths.

We should note that advocates of readability formulas have provided many counterarguments, but even a brief overview of the debate is beyond the scope of this paper. Our position is that, despite their shortcomings, readability formulas are useful, practical and objective predictors of text difficulty, but should be applied properly, in conjunction with other readability measures and with awareness of their limitations (see Chall and Dale 1995, for a more thorough discussion on the topic).

## 4. Collocations

### *4.1 Introduction*

"You shall know a word by the company it keeps!" (Firth, 1957)

This maxim, which sounds like a commandment, is one of the most cited in the literature regarding collocations. It is what one of the most prominent figures in British Linguistics, J.R. Firth, used in order to draw our attention to the fact that in natural language, words are not randomly combined, constrained only by syntax, but they have *preferences.* Firth coined the term collocations for such "habitual" word combinations. For example, in English it is proper to say "strong tea" but not "powerful tea". One can say "broad daylight" but not "bright daylight". Some forms are conventionally acceptable and others are not, and this appears to have arisen arbitrarily in the course of language evolution.

### *4.2 What do we mean by collocations?*

Simple as it may be, Firth's description of a collocation is quite vague. This in turn has given rise to a multitude of different definitions, based on the perspective of every researcher and *the particular applications for which collocations were to be used*. In other words, there is no general agreement on the definition of a collocation. This is mirrored in the large number of terms alternative to collocations that are used almost interchangeably, such as multi-word expressions (MWE), multi-word units (MWU), bigrams (or n-grams[12] in general), idioms and co-occurrences (Evert, 2004, p. 16). Consequently, we will describe the attributes of collocations, as generally recognised. Nevertheless, for the sake of argument, we will provide the definition given by Choueka (1988), which is quite satisfactory:

> "[A collocation is] a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning cannot be derived directly from the meaning or connotation of its components."

It is important to note that this definition assumes adjacency of words, but in linguistics a phrase can be a collocation even if it is not consecutive. Three additional criteria for a collocation to be called thus need to be mentioned (Manning and Schütze, 1999, p. 172):

- **Non-compositionality:** The meaning of a collocation cannot be directly derived from the meaning of its parts. Either the meaning is completely different from the free combination (like in the case of "kick the bucket") or there is an added element of meaning to the whole phrase that cannot be predicted from the parts.
- **Non-substitutability:** Components of a collocation cannot be substituted with similar ones and still hold its exact meaning. For example, in the case of "white wine", we cannot substitute white with yellow (and say "yellow wine"), even though yellow is a good approximation of the wine's true colour.
- **Non-modifiability:** Collocations cannot be freely modified with additional lexical material or through grammatical transformations. For example, the idiom "apple of your eye" cannot be modified into "apples of your eye" or "apple of your beautiful eye". If this is done, the collocation sounds unnatural.

[12] N-grams are groups of words. The number of words in an n-gram is given by the integer n, so for n=2 we have bigrams, for n=3 we have trigrams and so on.

There are minor divergences from these criteria, but overall they are typical of collocations. Furthermore, collocations exhibit the following commonalities (Smadja, 1993, p. 146): they are arbitrary (it is not clear why "pick up" means to "learn quickly"); they are domain-dependent ("interest rate", "stock market"); they are recurrent and cohesive lexical clusters: the presence of one of the collocation's components strongly suggests the rest of the collocation ("United" could imply "Kingdom" or "Nations").

Far from a complete overview of the ongoing discussion of what constitutes a collocation, the previous paragraphs hopefully served as a very brief and general introduction to the notion of collocation. As stated earlier, the meaning attached to the term collocation depends heavily on the particular application is being used for. As a result, in the confines of our research, *it is corpus frequency information and statistical association measures that decide what we class as a collocation and what we do not*. Thus, the collocation list we have extracted is just a "noisy" substitute for a list of manually validated collocations[13], but this was sufficient for the scope of our work. It is also important to note that no syntactic information was used when extracting collocations from the corpus. Finally, from our perspective, words collocate when they appear within a certain distance from each other, within a certain "collocational" span (Sinclair, 1991, p.175) measured by the number of intervening words.

### *4.3 Types of collocations*

A way of categorising collocations is by using the syntactic relations upon which they are based. With this as a criterion, the types of collocations often at focus in English are (Evert, 2004, p. 19):

1. Verb + noun (direct object), e.g. *commit suicide.*
2. Adjective + noun, e.g. *reckless abandon.*
3. Adverb + verb, e.g. *tacitly agree.*
4. Verb + predicative adjective, e.g. *keep sth handy.*
5. Verb + particle constructions, e.g. *bring sth up*
6. Verb + prepositional phrase, e.g. *set in motion.*

It is apparent that this typology[14] focuses at collocations that are word pairs, something not adequate for the scope of the present project. A more useful grouping of collocations is the one provided by Wermter and Hahn (2004), who divide collocations into three classes, based on varying degrees of semantic compositionality of the basic lexical entities involved. These classes are (taken from Wermter and Hahn, 2004):

- **Idiomatic Phrases**: In this case, *none* of the collocation components involved contribute to the overall meaning in a semantically transparent way, e.g. "kick the bucket". The meaning of the expression is metaphorical or figurative.
- **Support Verb Constructions/Narrow Collocations**: This second class contains expressions in which *at least one* component contributes to the overall meaning in a semantically transparent way and thus constitutes its semantic core, e.g. "little black book".
- **Fixed Phrases:** Here, *all* basic lexical meanings of the components involved contribute to the overall meaning in a semantically much more transparent way, e.g. "heart attack". Still, they are *not as completely compositional* as to classify them as free word combinations.

### *4.4 Applications of Collocations*

Collocations have found uses in a wide spectrum of applications. As stated in (Manning and Schutze, 1999, p. 142), they are used in:

[13] Fully automated means can only extract cooccurences (or collocation candidates) from a corpus. Human editing and evaluation are needed in order to identify which of the candidates are true collocations.

[14] A more comprehensive description of the syntactic patterns of collocations can be found in (Poulsen, 2005, p. 16).

- Natural language generation, in order to make the output sound as natural as possible and avoid mistakes like “take a decision”.
- Computational lexicography, for the automatic identification of collocations to be listed in a dictionary entry.
- Parsing, so that preference can be given to parses with natural collocations.
- Corpus linguistics research, e.g. to compare linguistic habits in languages or regions i.e. same word used in different contexts.

Additionally, collocations have applications in:

- Word sense disambiguation, by using context in order to identify the meaning of polysemous[15] words (Ide and Véronis, 1998).
- Language teaching, where student knowledge of collocations is essential for accuracy and fluency in the foreign language at focus (Nesselhauf, 2003, p. 223).
- Machine translation, since a high-quality translation is very dependent on machine-readable dictionaries of collocations and their translation equivalents (Smadja et al. 1996, p. 5).

### *4.5 Association measures*

Association measures are the criteria employed to decide whether any specific sequence of words qualifies as a collocation. In keeping with Wermter and Hahn (2004), such measures may be classified as:

- frequency-based measures (e.g., based on absolute and relative co-occurrence frequencies)
- information-theoretic measures (e.g., mutual information, entropy)
- statistical measures (e.g., chi-square, t-test, log-likelihood, Dice’s coefficient)

The following are the best known association measures used in collocation extraction (Evert, 2004, p. 21, McEnery and Wilson, 2001, p. 86-87):

- T-score
- Mutual Information
- Log-likelihood
- Dice coefficient
- Z-score

For our collocation extraction, we selected the *log-likelihood* measure, for the following reasons:

1. It is widely used in computational linguistics for collocation extraction and is generally regarded as the *de facto* standard.
2. In the evaluation of association measures provided by Evert (2004), log-likelihood consistently had some of the highest precision and recall scores[16].
3. It can be used for the extraction of multiword collocations.

## 5. Deriving a new readability measure

Current readability formulas use mainly intrinsic textual characteristics like word length and sentence length to determine the relative complexity of texts. This practice is founded on readability studies, but is also influenced by the fact that most formulas were used manually. Corpora and computers make the use of more general variables possible. By using more objective, language-level criteria like frequency data, extracted from corpora, the accuracy of

[15] A word is polysemous when it has more than one possible meaning.

[16] Since we talk about collocation extraction or retrieval, we can use precision and recall measures to evaluate the retrieval effectiveness. Recall measures the proportion of relevant information retrieved in response to a search procedure. Precision measures the proportion of retrieved items that are in fact relevant (Oakes, 1998, p.176).

readability formulas can be improved. The language-level variable that we propose is collocation frequency. A benchmark collocation frequency list will be derived from the British National Corpus.

The main aims of this project are to:

1. Detail the range of software tools available for use in collocation (and similar activities) and compare those used specifically for collocation extraction.
2. Extract a collocation frequency list from the BNC.
3. Create a readability formula that uses the classic readability variables along with collocation frequency data, to determine the relative complexity of a text.
4. Computerise the formula. Design and code a Java-based program that implements the aforementioned formula.
5. Compare the formula against classic readability formulas like Flesch-Kincaid, Flesch Reading Ease and the Gunning Fog Index.

Results from earlier readability research (Bormuth, 1969, cited by Klare, 1984, p. 687) show that more complex predictors with more variables, made available by using computers, are *not necessarily more powerful* than the simple ones. This may suggest that our insight is flawed but for the following reasons, we believe that it is sound:

1. The use of corpora as a basis for our readability measure makes possible the use of more accurate language-level variables that were not previously possible.
2. We do not seek to improve the accuracy of readability prediction by using complex combinations of multiple variables. Instead, we suggest the substitution of existing semantic difficulty indicators with a new, single indicator, which is collocation frequency.

Readability measures often rely on factors such as word frequency and average word length (the number of syllables divided by the number of words) and average sentence length (the number of words divided by the number of sentences). Some of these aspects serve as indicators of syntactic difficulty. For instance, the longer a sentence is, the heavier the memory and mental load it places on the reader (Bormuth, 1966, cited by DuBay, 2004, p.42). Thus, a longer sentence tends to be the more difficult than a shorter one. While not obvious, factors such as word frequency and word length are indicative of semantic difficulty. According to Zipf's Law (Zipf, 1949), it is easier to understand words that are used frequently in a language. Furthermore, the most frequently used words tend to become shorter.

Our approach differs from most readability measures in virtue of our proposed indicator of semantic difficulty. This is based on average collocation frequency (acf) and this is given by the following equation:

$[\, acf = (1/n_c)(\sum_{i=1}^{m} f_i * n_i) \,]$, where:

- $acf$ = average collocation frequency
- $n_c$ = total number of collocation occurrences in sample text
- $m$ = number of different types of collocations in sample text
- $f_i$ = frequency of collocation type $i$ in corpus
- $n_i$ = number of occurrences of collocation type $i$ in sample text

This formula will be combined with other factors that reflect the syntactic complexity features for sample texts, to provide an estimate of readability that we believe will prove more effective than conventional alternatives and which is grounded in a plausible theoretical approach to readability analysis.

**References**

**Burnard, L. (2000).** Reference Guide for the British National Corpus (World Edition). URL: http://www.natcorp.ox.ac.uk/docs/userManual/urg.pdf Last Accessed 27 Jul. 2006

**Choueka, Yaacov (1988).** Looking for needles in a haystack. Proceedings of RIAO '88, pp. 609–623.

**Dale, E., Chall, J.S. (1948).** A formula for predicting readability. Educational research bulletin, 27, pp. 11-20, 37-54.

**Dale, E., Chall, J.S. (1995).** Readability Revisited: The new Dale-Chall readability formula. Massachusetts: Brookline Books.

**DuBay, W.H. (2004).** The Principles of Readability. Costa Mesa, CA: Impact Information. URL: http://www.impact-information.com/impactinfo/readability02.pdf Last Accessed 13 Jul. 2006

**Evert, S. (2004).** The Statistics of Word Cooccurrences (Word Pairs and Collocations). Ph. D. dissertation, Universitat Stuttgart.

**Firth, J. R. (1957).** A synopsis of linguistic theory 1930–55. *Studies in linguistic analysis*. The Philological Society, Oxford, pp. 1–32.

**Gray, W. S., Leary, B. (1935).** What makes a book readable. Chicago: Chicago University Press.

**Hunston, S. (2002).** Corpora in Applied Linguistics. Cambridge: Cambridge University Press.

**Ide, N. and Véronis, J. (1998).** Word sense disambiguation: The state of the art. Computational Linguistics, 241, pp. 1-40.

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

**Klare, G.R. (1984).** Readability. Handbook of reading research, ed. P. D. Pearson. New York: Longman, pp. 681-744.

**Manning, C. D., Schütze, Hinrich (1999).** *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

**McEnery, A.M., Wilson, A. (2001).** Corpus Linguistics. Edinburgh: Edinburgh University Press.

**McLaughlin, G.H. (1969).** SMOG Grading – A New Readability Formula. Journal of Reading, 22, pp. 639-646.

**Nesselhauf, N. (2003).** The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. *Journal of Applied Linguistics.* Oxford University Press, 24(2), pp. 223 -242.

**Poulsen, S. (2005).** Collocations as a language resource (A functional and cognitive study in English phraseology). Ph. D. dissertation, University of Southern Denmark.

**Redish, J. (2000).** Readability Formulas Have Even More Limitations Than Klare Discusses. ACM Journal of Computer Documentation, 24(3), pp. 132-140.

**Seaton, J. (1975).** Readability tests for UK professional journals. Journal of Librarianship, 7, pp. 69-83.

**Smadja, Frank (1993).** Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1), pp. 143–177.

**Smadja, F., McKeown, K.R., Hatzivassiloglou, V. (1996).** Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1), pp. 1–38.

**Stenner, A. J., I. Horabin, Smith, D. R., Smith, R. (1988).** *The Lexile Framework*. Durham. NC: Metametrics, Inc.

**Tekfi, C. (1987).** Readability formulas: An overview. Journal of Documentation. 43(3), pp. 261-273.

**U.S. State Department (1998).** A Plain English Handbook: How to create clear SEC disclosure documents. URL: http://www.sec.gov/pdf/handbook.pdf Last Accessed 25 Jul. 2006.

**Wermter, J. and Hahn. U. (2004)** Collocation extraction based on modifiability statistics. *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland.

**Zipf, George Kingsley (1949).** *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA.