ABSTRACT
The present paper elaborates the nature of the Average Collocation Frequency and proposes specific roles for this measure. In addition to its potential as a factor in gauging readability, we propose that the ACF affords a practical measure for estimating the language complexity of texts for use in English language teaching. Particularly, in the context of English as a Second Language, collocational content represents a dimension of language sophistication that is especially challenging to learners. For this reason, the ACF has potential as a discriminator of texts that may allow us to match learning materials to differing levels of learner.

1. INTRODUCTION
In a previous paper (Anagnostou and Weir, 2007a), we introduced the concept of Average Collocation Frequency (ACF) and argued for its use in estimating the readability of texts. The ACF measure is built upon two elements; corpus-derived frequency lists and the linguistic phenomenon of collocations. To clarify this relationship we need to review the characteristics of existing readability formulae and collocations.

1.1 Readability formulae
Readability formulae work by using quantifiable textual aspects, in order to estimate the 'difficulty' inherent in that text. Conventionally, these tools focus on word length and sentence length, or variations on these constructs. Word length functions as an indicator of semantic difficulty, following Zipf’s Law (Zipf, 1949) which indicates that the most frequent words in a language, and hence the most familiar, tend to become shorter. In contrast, sentence length functions as an indicator of syntactic difficulty, since longer sentences impose heavier demands on memory and mental processing on the reader (Bormuth, 1966; cited by DuBay, 2004).

These aspects are well founded in readability studies (Dave and Chall, 1995), but suffer the limitations faced by early readability researchers. Most readability formulae were developed in the period from the 1940s to the 1970s, with an orientation towards manual use. Nowadays though, corpora consisting of a trillion words (Franz and Brants, 2006), computers capable of utilising such vast datasets and a whole ecosystem of software tools for textual manipulation are available, removing the limitations that shaped the initial stages of readability work. Such technologies make exploring the use of more general, language-level variables possible.

Such ‘newer’ factors may include word frequency. The frequency of words, as derived from large reference corpora, reflects a viable factor in estimating readability since more common words are likely to be familiar to more readers. Thereby, a text composed mainly of highly common words is likely to prove more readable (more comprehensible). In spite of the plausibility of employing word frequency in readability measures, there are currently few examples (cf. Weir and Ritchie, 2007; Stenner et al., 1988).

Our own approach to readability estimation takes a similar line of reasoning and assumes that including objective language-level criteria, such as frequency lists extracted from corpora, can improve the accuracy of readability formulae. The specific additional factor that we recommend is based not upon word frequency, but on the frequency of word collocations.
1.2 Collocations

In natural language, words are not combined randomly. Although constrained by grammar and syntax, words also have ‘preferences’. Firth (1957) named such ‘preferential’ word combinations ‘collocations’. For instance, 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.

There is little consensus on what constitutes a collocation and most researchers adopt a definition that best fits their perspectives. In the context of our research, we regard word combinations as collocations on the basis of statistical association measures. Such measures determine collocations by estimating the strength of ‘bond’ between different words.

Some linguists argue that collocations rather than individual words are the fundamental building blocks of a language. Regardless of whether this holds true or not, with collocations the ‘whole is greater than the sum of its parts’. This indicates that the meaning of a collocation cannot be derived directly from the meaning of its parts. Either the meaning is completely different from the free combination of its constituent words - as 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 - a characteristic also known as non-compositionality (Manning and Schütze, 1999).

This complex meaning in collocations constitutes a layer of semantics that is not currently considered by existing readability measures. This led us to propose collocational components as a factor in gauging readability. Essentially, we should be able to ascertain a more accurate estimate for the semantic difficulty of a text by including its collocations and not only individual words.

2. DERIVING THE AVERAGE COLLOCATION FREQUENCY

In order to arrive at a practical application for the semantic contribution of collocations we used the British National Corpus (BNC) as a reference corpus. From this corpus we extracted, using various tools and techniques, collocation frequency lists. A review of such collocation extraction tools is given in Anagnostou and Weir (2007b).

The process we adopted for extracting collocations was statistical in nature, as opposed to syntactical, and was based on association measures to identify co-occurrences (or collocation candidates). Such measures determine the strength of bond between contiguous words. Higher scores for an association measure, indicate greater probability that the co-occurrence extracted is a collocation.

Collocation frequency lists (CFLs) are similar in nature to word frequency lists (WFLs), in which the two main fields are the type of word and its frequency in the corpus - either as a percentage (relative frequency) or as number of occurrences (absolute frequency). The key difference in CFLs, is that we measure the frequency of collocations rather than individual words.

We propose that the frequency of a particular collocation in the CFL is an indicator of its semantic opacity, or simply put, its comprehensibility. This is a logical extension of Zipf’s Law into the realm of multiword expressions: the higher the frequency of occurrence for a collocation, the easier it will be to understand.

From this standpoint, we formulated the Average Collocation Frequency. This is a factor of semantic difficulty that utilises CFLs and as such is a departure from the philosophy of most readability formulae, which use a variant of word length to determine semantic difficulty. The ACF is given by the following equation:

\[
ACF = \frac{1}{n_c} \left( \sum_{i=1}^{n} f_i \cdot n_i \right) = f_1 \cdot \frac{n_1}{n_c} + f_2 \cdot \frac{n_2}{n_c} + \cdots + f_m \cdot \frac{n_m}{n_c}
\]

Where:

\[acf\] = average collocation frequency
\[n_c\] = total number of collocation occurrences in considered text
\[m\] = number of different types of collocations in considered text
\[f_i\] = frequency of collocation type \(i\) in reference corpus
\[n_i\] = number of occurrences of collocation type \(i\) in sample text
The variable $m$ is the number of different types of collocation in the considered text. Note that by ‘types of collocation’ we do not mean, for example, how many idiomatic phrases (Wermter and Hahn, 2004) or verb+noun constructs (Evert, 2004) are present. Rather, the focus in this context is on how many different instances of collocations exist in the text (see Figure 1 for an example and further clarification).

The ACF combines the ‘absolute’ weight of a collocational token in semantic opacity (based on a CFL) along with the ‘relative’ weight that it has inside the text (frequency of a collocation in the considered text). The ‘absolute’ weight is represented by the $\phi$, which is the frequency of collocation type $i$ (where $1 < i < m$) in the BNC - or whichever reference corpus is being used for the extraction of the CFL. The CFLs produced from the BNC are the main input to the ACF and they constitute the most crucial component of the measure, since they serve as the comparator for any considered text, and determine the ACF’s accuracy. The ‘relative’ weight of a collocation is represented by the following fraction, $n_i/n_c$, where $n_i$ is the number of times a collocation of type $i$ appears in the sample text and $n_c$ is the total number of collocational tokens in the same text. The aforementioned fraction appears when expanding the summation in calculating the ACF, as depicted in the extended form of the equation and in Figure 2.

Consider the following scenario, using the BNC Baby (approximately five million words) as the reference corpus: The collocation ‘manna from heaven’ appears once in the BNC Baby, thus it has a frequency of 0.2 occurrences per million words (pmw). This is a rare collocation, therefore it is difficult and its impact on the semantic opacity of a passage that includes it will be great. Furthermore, we take the view that the semantic impact of the collocation is increased if it appears more than once in a text. In other words, the higher the frequency of occurrence of a difficult collocation in a sample text, the harder understanding the text is going to be. Any measure of semantic difficulty based upon collocations needs to accommodate this fact, and this is the role that the ‘relative’ weight of a collocation plays in the ACF.

Figures 1 and 2 illustrate an example of calculating the ACF for a text in which three different types ($m=3$) of collocations have been identified (all values are hypothetical). Column $n_i$ includes how many times each collocation appears in the sample text. By adding the cells in this column we can calculate $n_c$, or the total number of collocation occurrences in the sample text. The cells in column $\phi_i$ are populated with the frequency of each collocation in the reference corpus (CFL data), which, in this case, is measured as occurrences per million words (pmw). Figure 1 provides all the data needed to proceed to the calculation of the ACF (Figure 2).

<table>
<thead>
<tr>
<th>Collocations found in the sample text</th>
<th>$n_i$</th>
<th>$\phi_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>put up with</td>
<td>2</td>
<td>60 pmw</td>
</tr>
<tr>
<td>kick the bucket</td>
<td>1</td>
<td>22 pmw</td>
</tr>
<tr>
<td>pull yourself together</td>
<td>4</td>
<td>46 pmw</td>
</tr>
</tbody>
</table>

Total number of collocation occurrences in sample text = $n = n_1 + n_2 + n_3 = 2 + 1 + 4 = 7$

Figure 1. Factoring Collocations in the ACF
3. APPLICATIONS OF THE ACF

Our primary intended application for the ACF is in readability formulae. As discussed in the previous sections of this paper, existing formulae do not include variables that tap the additional levels of meaning embodied in collocations and do not take advantage of the frequency data now available. The ACF should serve as a component in a readability formula and thereby enhance the accuracy and relevance of that readability measure.

For example, a new readability formula incorporating the ACF could have the following form:

\[ RF = a \cdot ACF + b \cdot WDI + c \cdot SDI + d \]

Where:

- RF = readability formula (arbitrary title)
- ACF = Average Collocation Frequency (semantic difficulty indicator on the level of collocations)
- WDI = Word Difficulty Indicator (commonly word length)
- SDI = Syntactic Difficulty Indicator (commonly sentence length)
- \(a, b, c, d\) = coefficients

The addition of the ACF is the main difference from most classical readability formulae and, by accounting for the collocational dimension of semantics, promises a step forward in the accuracy of such tools. The WDI and SDI components are generic terms representing the inclusion of difficulty indicators related to word content and syntax, respectively. In order to complete the formula, its coefficients need to be calculated through regression analysis. Anagnostou and Weir (2007a) provide more detailed on the steps needed to construct a readability formula. The 'dummy' formula above is a linear function of its variables, but there are alternative formats as well. Bormuth (1969, cited by Klare, 1984) argued that curvilinear formulae are more accurate in higher grade levels, whereas McLaughlin (1969) suggested that the components of a formula should be multiplied instead of added.
We have argued that the ACF is a plausible ingredient in a readability measure. Furthermore, given the semantic complexity of many collocations, in that the meaning of the whole is not directly derivable from the meaning of the component terms, we consider collocations to be a significant hurdle to learners of English. This will be especially so for ESL learners, since they will lack the exposure to many collocations that is available to native learners of English.

So, in addition to its role as a factor in readability estimation, we regard the ACF as a potent measure in its own right and envisage several specific applications in parallel to its use for readability. Recall that the ACF takes account of the 'commonality' of collocations (by reference to the frequency of occurrence of collocations in the BNC). Thereby, through the ACF, we gain a comparative measure of texts that allows us to rank such texts by their degree of collocational complexity. Such a ranking has potential as a basis for grading English language learning materials, both for native learners and second language learners of English. In addition, the ACF affords a means of matching the collocational content of learning materials to the learning audience (cf. Campbell & Weir, 2007).

Since second language learners are more likely to lack exposure to the ordinary application of collocations than native speakers of English, we consider that the ACF will yield greatest benefit in this context. Managed exposure to collocations that increase in complexity (i.e., are less frequent in their usage), promises to enhance the 'naturalness' of any second language learner's facility with English language use.

### 4. SUMMARY AND CONCLUSIONS

The ACF aims to quantify the impact of collocations on semantic opacity. As such, it holds potential and opens up a new path of research into a layer of semantics currently unaccounted for by traditional complexity measures. Of course, some of our perspectives on collocational semantic impact may require further defense.

For example, the ACF assumes that the repetition of a difficult collocation in a text further obfuscates its meaning and makes it harder to understand. While we regard this as intuitive, it is possible that on occasions, the repetition of a collocation might lessen the effect it has on semantic opacity because the reader might be able to deduce its meaning (or part thereof) from context. In this case, the more often the collocation appears, the higher the likelihood that its semantic opacity is reduced.

While this may appear to conflict with the basis of the Average Collocation Frequency, we suggest that such cases are exceptions rather than the norm. Ultimately, we may wish to accommodate such a possibility by gradually reducing the weight assigned to any repeated collocation. This is simply a variant that requires the empirical validation that remains for us to address.

A further counter may derive from earlier readability research to the effect that more complex predictors with more variables are not necessarily more powerful than the simple ones (Bormuth, 1969, cited by Klare, 1984). In fact, as argued in Anagnostou and Weir (2007a), our approach does not involve using complex combinations of multiple variables. Instead, we are proposing a simple indicator of semantic difficulty that embraces the wealth of meaning contained in collocations. Furthermore, we stand to benefit from corpus data that were non-existent when the aforementioned research was conducted and such data makes possible more accurate, language-level readability indicators, such as the ACF.

The theoretical underpinnings of the ACF lie in research into readability, corpus linguistics and collocations. Further peer review and testing are essential to ensure the measure’s validity. A software application that implements the ACF is currently in development. This will facilitate the required empirical testing.

### REFERENCES


