Future Internet **2012**, *4*, 238-252; doi:10.3390/fi4010238

OPEN ACCESS

future internet

ISSN 1999-5903

www.mdpi.com/journal/futureinternet

Article

Readability and the Web

Ludger Martin [1],* and Thomas Gottron [2]

[1] Institute of Computer Science, Johannes Gutenberg University Mainz, Mainz 55128, Germany

[2] Institute for Web Science and Technologies, Universität Koblenz-Landau, Koblenz 56070, Germany;
E-Mail:gottron@uni-koblenz.de

* Author to whom correspondence should be addressed; E-Mail: martin@informatik.uni-mainz.de.

Received: 20 December 2011; in revised form: 7 February 2012 / Accepted: 5 March 2012 / Published: 12 March 2012

Abstract: Readability indices measure how easy or difficult it is to read and comprehend a text. In this paper we look at the relation between readability indices and web documents from two different perspectives. On the one hand we analyse how to reliably measure the readability of web documents by applying content extraction techniques and incorporating a bias correction. On the other hand we investigate how web based corpus statistics can be used to measure readability in a novel and language independent way.

Keywords: web document readability; content extraction; corpus statistics

1. Introduction

Analysing a text for its readability, *i.e.*, the ease to read and comprehend the text, has a long tradition in literature. Since the nineteenth century, researchers in linguistics and literature science have been concerned with the question of how to measure the readability of a text. In recent years, the findings of this established field of research have found their application on web documents as well. Here, the notion of the readability of a document can serve several purposes. Readability metrics have been employed to assess the usability of web sites [1], as a static quality metric for ranking web search results [2] or to filter documents which match a user's reading ability [3–6].

However, the differences between classical print media and the web are often neglected in this transfer of readability metrics to the web. The differences of these media are twofold. On the one hand, text in the web is presented differently than in a printed book or magazine. This is motivated by technical restrictions on one or the other side, *i.e.*, a fixed page size *vs.* a scrollable screen viewport or the navigation via a table of contents and a look-up index with page numbers *vs.* the navigation via hyperlink structures and navigation menus. On the other hand, users consume text contents on the web differently compared to classical print media. Also here, the differences arise from usage patterns supported by the technologies, e.g., scanning and selective reading *vs.* linear and complete reading. In particular, user do not read all the text in a web document, but concentrate on the actually relevant parts while ignoring additional contents, such as navigation menus or advertisements.

Given these differences in the presentation and consumption of text, the application of readability metrics for the purpose of web documents needs to be reconsidered. In this paper we address this topic under several aspects. First we consider the problem of text noise in web documents. This noise represents additional texts in a document, that are typically not part of the main content and are perceived differently by a user. Obviously this noise should not be considered when determining the readability of a web document. One way to eliminate the noise is to provide hand crafted filters for cleaning the documents of a particular web site, which are all based on a common template. This approach requires to track changes in the templates [7] and to adapt the filters. Further, implementing hand crafted filters does not scale for an arbitrarily large number of web sites.

Thus, here we analyse and compare generic methods for removing the text noise in web document, so called content extraction (CE) algorithms. CE algorithms are generally applicable to all web documents and use document structure, text density or layout features to identify the main text content. As all state-of-the-art CE algorithms are based on heuristics and typically are not perfect, we look in a second step at the bias these methods introduce in the computation of readability metrics. As this bias is systematic we finally propose corrective measures to counterbalance the bias. Finally, we investigate the potential of using the web itself to define metrics for document readability.

Altogether in this paper we make several contributions.

- We bring together the works of readability assessment on web documents with content extraction techniques.
- We qualitatively evaluate the impact of general purpose content extraction methods on the estimation of the readability of a web document.
- We propose a bias correction depending on the CE methods which leads to improvements in readability estimation.
- We investigate the potential of designing domain specific readability metrics by incorporating web based reference corpora.

The rest of the paper is structured as follows: In Section 2 we introduce some of the more common readability formulae. We present related work in Section 3, with particular focus on the application of readability metrics on web documents and content extraction techniques. In Section 4 we consider how readability can be determined accurately for web documents. Section 5 investigates the potential to use web resources to compute a readability metric. Finally, we conclude the paper with an outlook at future work.

2. Readability Formulae

A readability index is a measure to express the complexity of written text. Quite often they are based on simple features, such as sentence and word length, and indicate how easy it is to read and comprehend a text. While there is a wide variety of readability indices covered in the related literature, we focus here on three well established methods: Flesch Reading Ease, the SMOG grading index and the Gunning fog index.

The *Flesch Reading Ease (FRE)* index [8] is a long established index in this context. The score of FRE typically ranges between 0 and 100 [9]. A higher score indicates a text that is easier to read and comprehend. For instance, a text with a score between 100 and 90 should be understandable for 11-year old students, while a score lower than 30 requires the reader to be at the level of a college graduate. Let us denote the total number of syllables in a text with y, the number of words with w and let s be the number of sentences, then the FRE index r_{FRE} is defined by:

$$r_{\text{FRE}} = 206.835 - 84.6 \cdot \frac{y}{w} - 1.015 \cdot \frac{w}{s} \tag{1}$$

McLaughlin [10] introduced a different parameter in his readability formula: the number of polysyllables. A polysyllable is a word made of three or more syllables. If we denote the number of polysyllables in a text with p, and use s again for the number of sentences, then the *SMOG grading* index r_{SMOG} is defined as:

$$r_{\text{SMOG}} = 1.043 \cdot \sqrt{\frac{p}{s}} + 3.1291 \tag{2}$$

The SMOG grading index indicates the educational level required to comprehend a text, *i.e.*, the years of school education, required to understand it. For instance, a SMOG reading index value of 5 indicates that a text can be understood after five years of school education. To calculate the index of a large text, McLaughlin stated that it is sufficient to use three text samples of 10 sentences each.

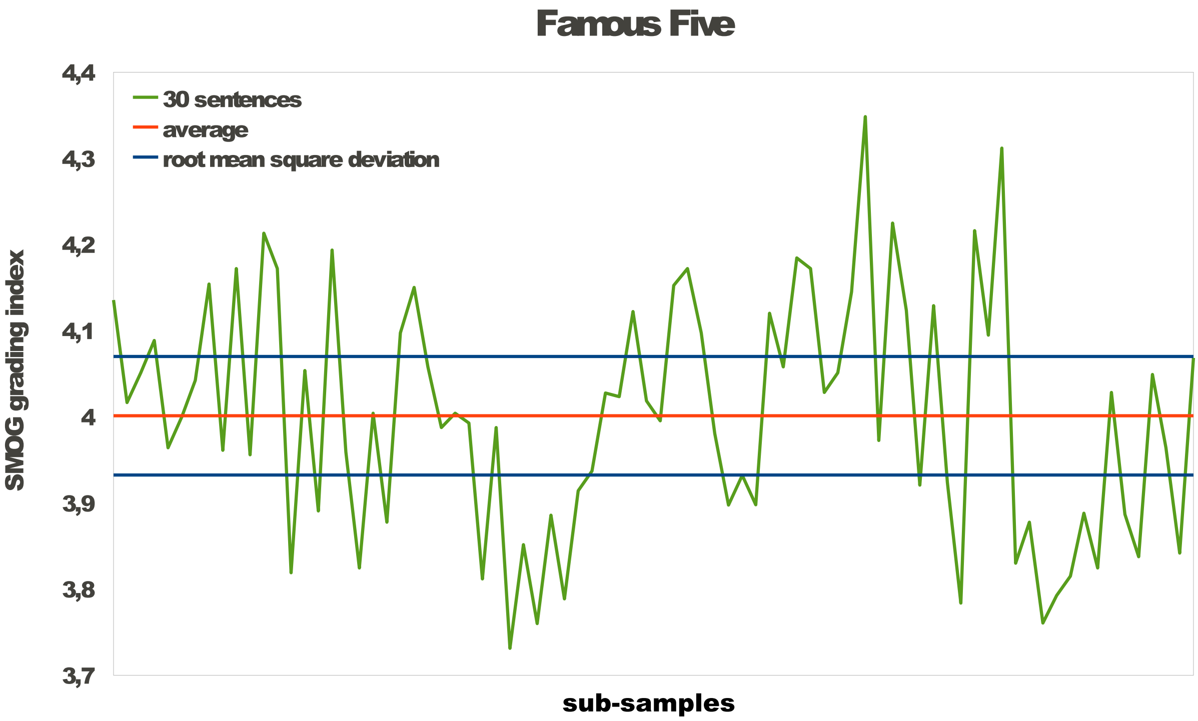
Another metric with a similar intention is the *Gunning fog* index [11]. Comparable to SMOG, also this index estimates the years of education required to understand a given text. To calculate the Gunning fog index, a passage of around 100 words needs to be analysed. Polysyllables are considered in the Gunning fog index, too, but only those which are not proper nouns, compound words, *etc*. The Gunning fog index r_{FOG} is defined as:

$$r_{\text{FOG}} = 0.4 \cdot \left(\frac{w}{s} + 100 \cdot \frac{p}{w}\right) \tag{3}$$

The original Gunning fog formula was based on clauses and not on the number of sentences s. This rendered the Gunning fog too difficult to be calculated automatically. Thus, by now, that the formulation presented in Equation (3) is generally recommended.

Both, the Gunning fog and the SMOG grading index use only small samples of a text to calculate the a readability score. This was originally simply motivated by practical reasons. Both formulae are relatively old. The Gunning fog index was published in 1952, the SMOG grading index in 1969. At this time the formulae needed to be calculated by hand. With the rise of electronic means to process texts both formulae have been implemented for automatic computation. Hence, their computation can easily be extended to a full text even for entire books. One question of interest is, whether there is a difference when considering the full text rather than just some parts and samples of a text. The chart in Figure 1 shows the development of the SMOG grading index across sub-samples of 30 sentences over a full text (in this case a novel for children).

Figure 1. Readability over sub-samples of a longer text.



The average of the SMOG grading index over all sub-samples is 4.00 and the SMOG grading index calculated with the whole text is 4.01. This difference can be neglected. The minimal index is 3.73 and the maximal index is 4.35. If one of these parts is selected randomly the readability might be miscalculated to the extent of half a school year.

The SMOG grading index and the Gunning fog index both promise to calculate the years of school education. Table 1 shows the SMOG grading and Gunning fog index of four texts. The table also shows the grade differences $\Delta = ||r_{\text{SMOG}} - r_{\text{FOG}}||$. The difference ranges from 2.33 up to 4.35 years of school education.

Table 1. Absolute difference Δ between SMOG and Gunning fog values for selected documents.

Document	**SMOG**	**Gunning fog**	**Δ**
Enid Blyton: Famouse Five—Five Are Together Again	4.01	6.34	2.33
William Shakespeare: Romeo and Juliet	4.05	7.04	2.99
Bible: Genesis 1–50	4.08	7.39	3.31
Arthur Conan Doyle: The Hound of the Baskervilles	4.27	8.62	4.35

This fluctuation in the results as well as the need to avoid a sampling bias motivated us to compute readability scores always on the full text we were considering.

While FRE and SMOG were defined for the English language, readability indices have also been developed for other languages. Amstad [12] describes a variation of FRE with adapted weights for German:

$$r_{\text{FRE}}^{\text{DE}} = 80 - 58.5 \cdot \frac{y}{w} - \frac{w}{s} \tag{4}$$

The *Wiener Sachtextformel*, instead was directly developed to describe the ease of readability for German texts [13]. It exists in different versions to cope with discontinuities and non-linear developments of the difficulty in reading German texts.

For some other languages, there are no readability indices or at least none are well established. To overcome this problem Tanguy and Tulechki [14] looked at an approach for identifying linguistic characteristics that might be most suitable to describe sentence complexity. Their approach was applied to French texts. However, they did not develop any formula nor did they investigate if the found features really correlate with readability.

3. Related Work

While the readability indices we recalled in Section 2 were originally developed for text documents, they are nowadays applied to web documents, too. Typical applications in this context are to assess how comprehensible legal statements are [1] or to let users select documents most appropriate to their educational level. The latter application can be subdivided further into ranking web search results according to their readability [2] or to filter documents that do not comply with the reading ability of a given user [3,4].

So far, little work consider the differences in the calculation of readability metrics for web documents. Petersen and Ostendorf [15] considered in particular the distinction between web documents that contain little text and for which it consequently is not suitable to compute a readability score at all. This distinction is based on a supervised learning approach using very common words as feature set for the documents. Few systems address the problem of identifying the actual main content in a web document when computing its readability. Nonaka *et al*. [2] present a domain and language dependent approach to detect the particular structure of how-to manuals. The Read-X system [4,16] instead, uses a heuristic content extraction tool to improve the estimation of readability. The extraction module in Read-X is based on a screen scraper software library supporting the development of hand crafted extraction filters for template based documents. However, so far no system involves state-of-the-art generic content extraction algorithms.

Generic content extraction algorithms are covered in various publications. Typically the algorithms are based on heuristics and follow some assumptions about the shape, structure and form of the main content. A thorough evaluation of different approaches is presented in [17]. In this comparison the Document Slope Curve filter [18] showed good efficiency and effectiveness. Newer methods have advanced the accuracy (e.g., Content Code Blurring [19]), efficiency (e.g., the Density algorithm [20]) or addressed language specific scenarios (e.g., the DANA approach for arabian languages [21]). All modern CE methods are highly efficient, are capable of processing between 10 and 20 MB of web documents data per second on commodity hardware and are, thus, suitable for on-the-fly application for cleaning web documents from noise. State-of-the-art extraction algorithms achieve an extraction performance with an average F1 score between 0.86 and 0.96, depending on the application scenario. A detailed analysis also showed that some methods are more biased towards a higher precision, while others favour a high recall in extracting the main content [17] .

In [22] we looked at the influence of noise in web documents—such as navigation menus, related links lists, headers, footers, disclaimers—on the automatically determined readability score. We found that content extraction algorithms like the Document Slope Curve filter [18] or Content Code Blurring [19] led to much better estimates of the readability of the document's actual main content.

Yan *et al.* [23] investigate domain specific readability. They explain that domain specific texts have technical or professional terms. These terms cannot be measured by traditional syllable counting like SMOG or FRE. Common words will become technical terms in domain specific texts. They also propose to compute a relative readability score rather an absolute grade level metrics. They present several complex formulae to calculate their concept-based document readability. They include calculations on word level, consider a given knowledge base and document scope, and also calculate general words which are out of a common word list. Yan *et al.* state that traditional readability formulae are oversimplified. The drawback of their approach instead lies in the need for a domain expert to formulate the required knowledge base. A general purpose formula that automatically determines the characteristics of a domain specific language would be favourable.

There are several approaches looking at alternative features for determining the readability of text. For instance, Rosa and Eskenazi [24] consider word complexity. In the context of computing the readability of a text written in a language that is not the reader's native language, they try to determine factors that make an individual word easier to learn. One such factor is word complexity. It can be measured by the word's grapheme to phoneme ratio and the number of meanings a word has. Another approach in a similar study for French as a foreign language [25] focusses on multi-word expressions (MWE) as the basis for a formula for measuring readability. The score of the formula is related to the *Common European Framework of Reference for Languages*. The most important variables are the proportion of the nominal MWEs to the number of words and the mean size of the nominal MWEs in the text. The first is significantly related to the difficulty of the text. The option of using word frequencies to determine readability is raised by Weir and Ritchie [26]. In Section 5.1 we follow this line of thought, and investigate a similar feature to determine the readability of a text.

4. Readability of Web Documents

Since the rise of the World Wide Web, more and more texts appear online and as part of web sites. Naturally, producers as well as consumers of online texts are interested in the readability of online documents. This has led to the application of readability formulae to HTML documents [1]. The typical approach here was to simply take a full HTML document, strip off all the markup and parse the remainder of the document through a readability formula. The result of the formula is then interpreted as the readability of the document.

While this is a straightforward solution to the problem, there is a conceptual mistake in the approach. As we have already mentioned before, web documents are designed and consumed differently than classical printed documents. In particular user do not read a document entirely, but rather scan the web page first to determine, where the main content is located and whether this content is relevant and of interest to them. Once the user has identified relevant information, they read the document selectively and focus on those parts comprising the main content. Other, additional contents, such as navigation menus, related links list, legal disclaimers, advertisements or header and footer elements with text are typically ignored when reading a document. So, essentially, as these text contents are not actually read, they should be ignored in the computation of readability metrics.

4.1. Content Extraction

Content Extraction (CE) is the process of determining those parts of an HTML document which represent its main text content. Hence, it is a suitable solution to address the problem described above. A qualitative evaluation of several CE approaches in [17] showed that modern methods demonstrate a very good performance in terms of accuracy. State-of-the-art methods achieve F1 scores of 0.96. However, CE methods are typically not capable of perfectly extracting the main content. Some methods tend to be too restrictive and discard some parts of the main content during the extraction process, others instead are too lax and extract also additional content. This bias in the methods needs to be taken into account when designing applications incorporating CE.

In the context of this paper we apply two established and well performing methods: *Adapted Content Code Blurring (ACCB)* [19] and *Document Slope Curves (DSC)* [18]. The details of these algorithms are beyond the scope of this paper and can be found in the original publications.

4.2. Experimental Setup

We base our work on data and initial experiments we conducted in [22]. To evaluate and quantify the impact of noise in web documents we crawled 1114 web documents from five different web sources. All sources provided English news articles with a text based main content. The length of these main content typically ranged between 300 to 1100 words, with a very few outliers of significantly shorter or longer documents.

To provide a gold standard we manually determined the actual main content in each of the documents and calculated FRE and SMOG on these texts. As baseline, we applied the same readability metrics of the full documents including all the text noise, such as navigation menus, headers, footers, advertisements, legal disclaimers, *etc*. Finally, we automatically cleaned the documents from additional content by using the ACCB and DSC filters and evaluated the readability of the remainder of the document.

For the computation of the readability metrics we determined sentence boundaries based on end-of-sentence characters and text structure. We paid attention to the context of end-of-sentence characters, e.g., by requesting a subsequent white space character and by checking against a list of common abbreviations for not detecting a premature end of a sentence. Additionally we considered a sentence to start at the beginning of each paragraph and to stop at the end of a paragraph. We tokenized sentences into words at white space characters and further special characters, like colon, comma, quotation marks, *etc*. We did not employ compound splitters or other more sophisticated methods. For the decomposition of words into syllables we relied on the hyphenation of the LaTeX package, which can easily be incorporated into other programs. This provided us all the features necessary for the computation of SMOG and FRE.

4.3. Results

Table 2 shows the values we obtained for the SMOG index, when calculating it on our gold standard (the actual main content), on the full document and after having cleaned the documents using ACCB and DSC. The values indicate clearly that employing CE (columns ACCB and DSC) in the course of determining readability provides more accurate results. However, the aggregated values do not show the variations and fluctuations of the values on individual documents, which brought us to analyse the results in more detail.

Table 2. Average SMOG index value for web documents based on the actual article text, the full text and after cleaning the document using the content extraction methods ACCB and DSC.

Source	**Number of documents**	**SMOG**			
		Gold standard	**Full**	**ACCB**	**DSC**
BBC News	337	4.8323	4.0569	4.9360	4.8052
The Economist	53	5.0578	4.2486	5.1433	5.0835
Herald Tribune	300	5.0477	4.0891	5.0650	5.0412
MSNBC News	197	4.8949	4.4675	4.9050	4.8491
Yahoo News	227	4.9416	4.2063	4.7563	4.7670
Total	1114	4.9344	4.1793	4.9385	4.8820

We measured the correlation of the readability of the actual hand cleaned main content with the readability values obtained for the full document and the ones for the automatically cleaned documents. As shown in Table 3 it turned out that the actual readability of a document and the values on the full document are at best weakly correlated, while the cleaned documents show a relatively good correlation. Further, we looked at the mean square error (MSE) of the data series to measure the local deviations. Also here employing ACCB and DSC leads to far better values.

Table 3. Correlation and MSE in measuring readability for different choices of text samples.

Choice of text sample	**Correlation**	**MSE**
Full document	0.3220	0.6624
ACCB	0.8406	0.0277
DSC	0.8612	0.0277

4.4. Bias Correction in Readability on the Web

The good correlation values obtained when cleaning the documents on the basis of the ACCB and DSC content extraction filters indicate a linear relation between the readability value on the actual content and the index as it is computed on the automatically created extract. However, looking at the individual values in Table 2, ACCB, for instance, tends to overestimate the level of difficulty for readability.

Our hypothesis is that these deviations are systematic and are caused by the bias of CE algorithms to achieve either a better precision or better recall when determining the main content. If this hypothesis is correct, we can correct the deviations by learning a functional relation between the true readability values obtained on the gold standard and the estimates obtained via the CE methods.

To learn such a functional relation, we sampled a random subset of 100 documents from our test corpus to train a linear regression model on the data. Using the results of this model we can correct the bias in the computation of the readability. To furthermore check if such a bias correction in readability would be sufficient to replace CE methods entirely, we also trained a model on the SMOG index values of the full documents. This lead to the following adjusted bias-correction (bc) formulae for SMOG on different text samples:

$$\mathit{SMOG}^{bc}_{full} = 0.9865 \cdot \mathit{SMOG}_{full} + 0.8241 \tag{5}$$

$$\mathit{SMOG}^{bc}_{ACCB} = 0.9284 \cdot \mathit{SMOG}_{ACCB} + 0.3573 \tag{6}$$

$$\mathit{SMOG}^{bc}_{BSC} = 0.9525 \cdot \mathit{SMOG}_{BSC} + 0.2718 \tag{7}$$

We evaluate the bias correction by computing again MSE for the adjusted values on the remaining 1.014 documents. The results in Table 4 show that the bias correction reduces MSE in all series. The biggest relative improvement for SMOG is obtained on the full documents. However, the best absolute value is achieved for the bias corrected SMOG index on the DSC filter. For the ACCB content extraction filter the improvements are still valid but do not reach the quality of DSC.

Table 4. MSE with and without bias correction.

Choice of text sample	No correction	Bias correction
Full document	0.6624	0.0833
ACCB	0.0277	0.0249
DSC	0.0277	0.0234

5. Exploiting Web Resources to Estimate Readability

The readability metrics SMOG, FRE, and Gunning fog have been designed for English. The adaptation of FRE for German required an adjustment of parameters, for most other languages there are no parameters for this metric. Furthermore, for some languages there are no readability metrics at all.

5.1. Web Metrics for Documents

Fortunately, linguists have observed that also other features correlate with the difficulty of reading and understanding a text. One such feature is how common the words in a text are [13]. A text is easier to comprehend if it contains mainly commonly used words and harder if it uses words that are not part of everyday language.

Frequency classes are one approach to measure how common a word is. The frequency class of a word quantifies how much less frequent a word is than the most frequent word of a language. Class c_0

is assigned to the most frequent word, class c_1 to all words that are at least half as frequent as the one in c_0. In turn, class c_2 contains all words that are at least half as frequent as the words in c_1, and so on. The class of a term t w.r.t. the most frequent term t_0 can be computed in a closed form:

$$c(t) = 0.5 - log\left(\frac{freq(t)}{freq(t_0)}\right) \tag{8}$$

To compute the term frequencies in a language it is necessary to have statistics over a large corpus. The largest electronically accessible corpus is provided by the web. The Wortschatz project of the University of Leipzig [27] has massively crawled large amounts of web documents in different languages. The project provides a SOAP interface to its database that allows for easy querying corpus statistics and also for directly obtaining term frequency classes for terms.

5.2. Analysis of Texts

We build an application that can take up any arbitrary text, tokenizes it into words and retrieves via the web service interface of the Wortschatz project the frequency class for each word. We then analysed how many distinct words are in every frequency class and obtain a distribution of terms to frequency classes.

We created a collection of publicly available texts and assigned each text into one of the classes of small children's literature, novels, scientific texts, news and philosophical manuscripts. For news we used a sub-sample of our dataset described above in 4.2, the philosophical texts were obtain from project Gutenberg [28]. As scientific text we used the full papers published at the 9th WWW conference and the children's literature and novels were text samples taken from recent books or web published short stories of contemporary authors. Table 5 lists how many documents are contained in each category.

Table 5. Classes of analysed texts.

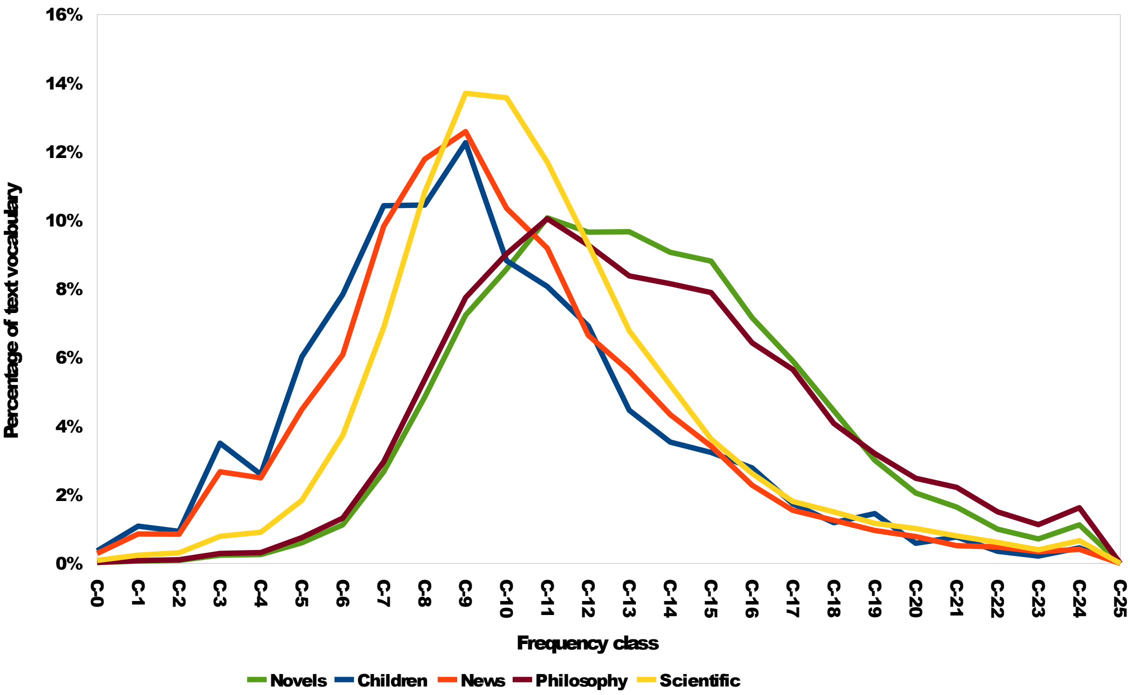
Text category	Number of documents
Children	19
Novels	14
Scientific	57
News	198
Philosophy	5

The graph in Figure 2 displays the distribution of terms into the frequency classes in the different text categories. The plots show that the distributions do correspond to the difficulty of the texts. However, there are several interesting aspects:

- Children's literature as well as news articles use few rare words. Looking into the texts, the rare words mainly corresponded to names of people or locations, such as towns, rivers, countries. While at first sight it might seem surprising that news and texts for children should be comparable in terms of readability, our above results also showed that news texts on average have a SMOG index of 5 which approximately corresponds to children having completed primary school.

- Relative to the other categories of texts, the very rare and very common words are under-represented in scientific texts. In the middle range frequency classes the WWW papers have a higher percentage of words, which can be explained with a scientific community having its particular but not too common language.
- Novels contain more rare words. In philosophical texts this observation is even stronger. For the novels this can be explained in a well elaborated text style, for the philosophical texts with a very high level of writing.

Figure 2. Distribution of terms in frequency classes for different types of texts.



Given the distribution of the terms we can estimate the expected frequency class $E(c(t))$ for a randomly chosen term from a text in a given category. The values in Table 6 show this expected frequency that corresponds to the intuitive order of perceived difficulty of the texts in a category.

Table 6. Expected frequency classes of analysed texts.

Text category	$E(c(t))$
Kids	10.58
Novels	12.79
Scientific	11.85
News	10.83
Philosophy	14.28

5.3. Frequency Classes and Readability

Given the observations in the last section, the question arises, whether it is possible to automatically estimate the readability of a given text based on its frequency classes. In the previous section we observed the expected frequency class $E(c(t))$ to correspond with an intuitive ranking of the readability of text categories. While above we looked at entire collections of texts from the same categories, we now change our focus to individual documents and see if we can observe a similar pattern.

To gain some first insights we chose four different texts. Table 7 shows the readability of these texts according to SMOG and Gunning fog. Additionally the table shows the expected frequency class. It can be seen, that on this document level the three metrics do not agree. The values do not imply the same ranking of the documents.

Table 7. SMOG, Gunning fog and expected frequency class for selected documents.

Document	**SMOG**	**Gunning fog**	$\boldsymbol{E(c(t))}$
Enid Blyton: Famouse Five—Five Are Together Again	4.01	6.34	12.86
William Shakespeare: Romeo and Juliet	4.05	7.04	13.65
Bible: Genesis 1–50	4.08	7.39	13.29
Arthur Conan Doyle: The Hound of the Baskervilles	4.27	8.62	12.95

In a second, more extensive experiment we computed FRE and the expected frequency class $E(c(t))$ for 60 documents of different readability. The readability of these documents ranged from an FRE index of 28.073 for the most difficult to a value of 88.499 for the easiest document. The value of $E(c(t))$, instead, was in the ranged between 8.904 and 14.745. The values of FRE and $E(c(t))$ showed a slight, but no significant Pearson correlation of 0.47. Thus, we can say that while the expected frequency class can give indications about the readability of a document, as an exclusive feature it is not sufficient to provide a precise analysis for a given text.

This is quite surprising, given that the expected frequency class provided a good insight into the readability of an entire collection of documents belonging to the same category. One possible explanation for this unexpected behaviour on individual documents is data sparsity. A single document might not contain enough different words to extract a sound model of the term distribution over frequency classes. Hence, it might be necessary to apply smoothing methods to the observed distribution in order to get more reliable results. Another approach that can be pursued in parallel is to improve the results by using more information about the frequency class distribution itself for the prediction of the readability. Such additional information could come in the form of estimating the variance or directly using the actual discrete distribution. Beyond feature engineering, it might additionally be necessary to adjust a non-linear function to derive the level of readability from the multi-dimensional input data obtained from the frequency class distribution. However, while we laid the foundation for this analysis, the concrete steps to be taken are left for future work.

6. Conclusions and Future Work

In this paper we analysed the relation between readability indices and the World Wide Web. We looked at the topic from two different angles: How to determine readability of documents on the web and on the potential of exploiting web resources to indicate parameters for new approaches to determine readability.

Concerning the application of readability measures on web documents we showed that the introduction of content extraction filters into the process leads to significantly improved estimates. Further, we developed bias adjustments for CE based SMOG and FRE indices that lead to still better estimates for the readability of web documents. Given that we focused in our analysis on news documents, it remains to investigate how the CE methods operate on other type of documents or documents with different levels of readability.

On the other hand we found indications that corpus statistics of the web can be exploited to obtain language independent measures for readability. We showed that the distribution of terms into frequency classes reproduces very nicely the intuitively perceived difficulty of text categories. However, when looking at the document level, the latter results require some further investigations. Predicting the readability for a single document simply based on the expected frequency class does not provide results of the desired quality yet. Overcoming data sparsity in single documents and using more characteristics and features of the frequency class distribution seems a promising approach here.

In future work we will address exactly this task of feature engineering on the frequency class distribution in individual documents. We are confident, that by applying smoothing techniques and identifying a set of suitable features it will be possible to estimate the readability also of individual documents based on the frequency classes of the contained terms. Once such a metric is established, an interesting question will be if the observations can be generalized to other languages, thereby providing a language independent readability metric.

Acknowledgements

The research leading to these results has received partial funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 257859, ROBUST.

References

1. Kienle, H.; Vasiliu, C. Evolution of Legal Statements on the Web. In *Proceedings of the 10th IEEE International Symposium on Web Site Evolution*, Beijing, China, 3–4 October 2008; pp. 73–82.
2. Nonaka, R.; Yumoto, T.; Nii, M.; Takahashi, Y. Finding How-to Information Web Pages and Their Ranking by Readability. In *Proceedings of the IADIS International Conference Internet Technologies and Society (ITS '10)*, Perth, Australia, 29 November 2010; pp. 155–163.
3. Lau, T.P.; King, I. Bilingual Web Page and Site Readability Assessment. In *Proceedings of the 15th international conference on World Wide Web (WWW '06)*, Edinburgh, UK, 22–26 May 2006; ACM: New York, NY, USA, 2006; pp. 993–994.

4. Miltsakaki, E.; Troutt, A. Real-Time Web Text Classification and Analysis of Reading Difficulty. In *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications (EANL '08)*, Columbus, OH, USA, June 2008; Association for Computational Linguistics: Stroudsburg, PA, USA, 2008; pp. 89–97.
5. Hussain, W.; Sohaib, O.; Ali, A. Improving web page readability by plain language. *IJCSI Int. J. Comput. Sci. Issues* **2011**, *8*, 315–319.
6. Collins-Thompson, K.; Callan, J. Information Retrieval for Language Tutoring: An Overview of the REAP Project. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04)*, Sheffield, UK, 25–29 July 2004; ACM: New York, NY, USA, 2004; pp. 544–545.
7. Gottron, T. Detecting Website Redesigns via Template Similarity on Streams of Documents. In *Proceedings of the 3rd International Conference on Internet Technologies and Applications (ITA '09)*, Wuhan, China, 18–20 August 2009.
8. Flesch, R. A new readability yardstick. *J. Appl. Psychol.* **1948**, *32*, 221–233.
9. From the formal definition it becomes obvious that FRE can also produce values out of the intended range, when applied to non standard texts.
10. McLaughlin, G.H. SMOG grading: A new readability formula. *J. Read.* **1969**, *12*, 639–646.
11. Gunning, R. *The Technique of Clear Writing*; McGraw-Hill International Book Co.: New York, NY, USA, 1952.
12. Amstad, T. *Wie Verständlich Sind Unsere Zeitungen?*; Dissertation, University Zürich: Zürich, Switzerland, 1978.
13. Köhler, R.; Altmann, G. Synergetische aspekte der linguistik. *Z. Sprachwiss.* **1986**, *5*, 253–265.
14. Tanguy, L.; Tulechki, N. Sentence Complexity in French: A Corpus-Based Approach. In *Proceedings of the 17th International Conference Intelligent Information Systems (IIS 09)*, Kraków, Poland, 16–18 July 2009; pp. 131–144.
15. Petersen, S.E.; Ostendorf, M. Assessing the Reading Level of Web Pages. In *Proceedings of the ICSLP 9th International Conference on Spoken Language Processing (INTERSPEECH '06)*, Pittsburgh, PA, USA, 17–21 September 2006; pp. 833–836.
16. Miltsakaki, E. Matching Readers' Preferences and Reading Skills With Appropriate Web Texts. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session (EACL '09)*, Athens, Greece, 30 March–3 April 2009; Association for Computational Linguistics: Stroudsburg, PA, USA, 2009; pp. 49–52.
17. Gottron, T. Evaluating Content Extraction on HTML Documents. In *Proceedings of the 2nd International Conference on Internet Technologies and Applications (ITA '07)*, Wrexham, North Wales, UK, 4–7 September 2007; pp. 123–132.
18. Pinto, D.; Branstein, M.; Coleman, R.; Croft, W.B.; King, M.; Li, W.; Wei, X. QuASM: A System for Question Answering Using Semi-Structured Data. In *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '02)*, Portland, OR, USA, 14–18 July 2002; ACM: New York, NY, USA, 2002; pp. 46–55.

19. Gottron, T. Content Code Blurring: A New Approach to Content Extraction. In *Proceedings of the 19th International Workshop on Database and Expert Systems Applications (DEXA '09)*, Turin, Italy, 1–5 September 2008; pp. 29–33.
20. Moreno, J.; Deschacht, K.; Moens, M. Language Independent Content Extraction From Web Pages. In *Proceeding of the 9th Dutch-Belgian Information Retrieval Workshop*, Enschede, The Netherlands, 2–3 February 2009; pp. 50–55.
21. Mohammadzadeh, H.; Gottron, T.; Schweiggert, F.; Nakhaeizadeh, G. A Fast and Accurate Approach for Main Content Extraction based on Character Encoding. In *Proccedings of the 8th Workshop on Text-based Information Retrieval (TIR '11)*, Toulouse, France, 29 August–2 September 2011; Unpublished work.
22. Gottron, T.; Martin, L. Estimating Web Site Readability Using Content Extraction. In *Proceedings of the 18th International World Wide Web Conference (WWW '09)*, Madrid, Spain, 20–24 April 2009; pp. 1169–1170.
23. Yan, X.; Song, D.; Li, X. Concept-Based Document Readability in Domain Specific Information Retrieval. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, Arlington, VA, USA, 6–11 November 2006; ACM: New York, NY, USA, 2006; pp. 540–549.
24. Rosa, K.D.; Eskenazi, M. Effect of Word Complexity on L2 Vocabulary Learning. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications (IUNLPBEA '11)*, Portland, OR, USA, 24 June 2004; Association for Computational Linguistics: Stroudsburg, PA, USA, 2011; pp. 76–80.
25. François, T.; Watrin, P. On the Contribution of MWE-based Features to a Readability Formula for French as a Foreign Language. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011 (RANLP '11)*, Hissar, Bulgaria, 12–14 September 2011; pp. 441–447.
26. Weir, G.R.S.; Ritchie, C. Estimating Readability with the Strathclyde Readability Measure. In In *Proceedings of the ICT in the Analysis, Teaching and Learning of Languages (ICTATLL'06)*, Glasgow, UK, 21–22 August 2006.
27. Quasthoff, U.; Richter, M.; Biemann, C. Corpus Portal for Search in Monolingual Corpora. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC '06)*, Genoa, Italy, 24–26 May 2006.
28. Project Gutenberg. Available online: http://www.gutenberg.org/ (accessed on 28 January 2011).