What can readability measures really tell us about text complexity?

Article · January 2012

CITATIONS
39

READS
1,571

4 authors:

Sanja Stajner
67 PUBLICATIONS **1,529** CITATIONS
SEE PROFILE

Richard Evans
University of Wolverhampton
48 PUBLICATIONS **855** CITATIONS
SEE PROFILE

Constantin Orasan
University of Surrey
158 PUBLICATIONS **2,053** CITATIONS
SEE PROFILE

Ruslan Mitkov
University of Wolverhampton
138 PUBLICATIONS **3,472** CITATIONS
SEE PROFILE

What Can Readability Measures Really Tell Us About Text Complexity?

Sanja Štajner, Richard Evans, Constantin Orăsan, and Ruslan Mitkov

Reseach Institute in Information and Language Processing
University of Wolverhampton
S.Stajner@wlv.ac.uk, R.J.Evans@wlv.ac.uk, C.Orasan@wlv.ac.uk, R.Mitkov@wlv.ac.uk

Abstract

This study presents the results of an initial phase of a project seeking to convert texts into a more accessible form for people with autism spectrum disorders by means of text simplification technologies. Random samples of Simple Wikipedia articles are compared with texts from News, Health, and Fiction genres using four standard readability indices (Kincaid, Flesch, Fog and SMOG) and sixteen linguistically motivated features. The comparison of readability indices across the four genres indicated that the Fiction genre was relatively easy whereas the News genre was relatively difficult to read. The correlation of four readability indices was measured, revealing that they are almost perfectly linearly correlated and that this correlation is not genre dependent. The correlation of the sixteen linguistic features to the readability indices was also measured. The results of these experiments indicate that some of the linguistic features are well correlated with the readability measures and that these correlations are genre dependent. The maximum correlation was observed for fiction.

Keywords: text simplification, readability, autism spectrum disorders

1. Introduction

Text simplification can be regarded as the process of converting input text into a more accessible form. The conversion process may be facilitated by research in various areas of NLP, including lexical simplification (Yatskar et al., 2010), anaphora resolution (Mitkov, 2002), word sense disambiguation (Escudero et al., 2000), syntactic simplification (Siddharthan, 2006; Evans, 2011), text summarisation (Orăsan and Hasler, 2007), or image retrieval (Bosma, 2005).

In the context of personalisable applications, it is necessary for systems not only to simplify text, but also to discriminate between material that should be simplified and material that should not be, for the benefit of a particular user. This discrimination can be realised by quantifying the difficulty of the material by means of various features of the text, and comparing those feature values with thresholds specified in user preferences.

The work described in this paper is part of an ongoing project that develops tools to help readers with autism spectrum disorders (ASD). One of the prerequisites for this research is to have a way to assess the difficulty of texts. A set of metrics is proposed with the aim of quantifying the difficulty of input documents with respect to their requirements. This set contains readability indices and metrics inspired by the needs of people with ASD. Documents from several genres are evaluated with regard to these metrics and the correlation between them is reported.

1.1. Requirements of Users with Autism Spectrum Disorders

This paper presents research undertaken in the initial phase of FIRST,[1] a project to develop language technology (LT) that will convert documents from various genres in Bulgarian, English, and Spanish into a more accessible form for readers with autism spectrum disorders (ASD).

ASD are defined as neurodevelopmental disorders characterised by qualitative impairment in communication and stereotyped repetitive behaviour. They are serious disabilities affecting approximately 60 people out of every 10 000 in the EU. People with ASD usually have language deficits with a life-long impact on their psychosocial functioning. These deficits are in the comprehension of speech and writing, including misinterpretation of figurative language and difficulty understanding complex instructions (Minshew and Goldstein, 1998). In many cases, people with ASD are unable to derive the gist of written documents (Nation et al., 2006; O'Connor and Klein, 2004; Frith and Snowling, 1983).

Written documents pose various obstacles to reading comprehension for readers with ASD. These include:

1. Ambiguity in meaning:
   (a) Figurative language such as metaphor and idioms,
   (b) Non-literal language such as sarcasm,
   (c) Semantically ambiguous words and phrases,
   (d) Highly specialised/technical words and phrases.
2. Structural complexity:
   (a) Morphologically, orthographically, and phonetically complex words,
   (b) Syntactically complex sentences,
   (c) Inconsistent document formatting.

A detailed study of user requirements derived from a focus group partially supported the initial hypothesis of their reading comprehension difficulties. The focus group made recommendations for the automatic simplification

[1] A Flexible Interactive Reading Support Tool (http://www.first-asd.eu).

of phenomena at various linguistic levels. This includes the automatic expansion and elaboration of acronyms and abbreviations (obstacle 1d); the replacement of ambiguous words/homographs by less ambiguous words (obstacle 1c); the substitution of anaphoric references by their antecedents, especially in the case of zero anaphora (obstacle 1c); the rewriting of long sentences as sequences of short sentences, the conversion of passive sentences into active sentences (obstacle 2b); and the translation of phraseological units such as collocations, idioms, and ironic/sarcastic statements into a more literal form (obstacles 1a and 1b).
In addition to the removal of obstacles to reading comprehension, recommendations were also made for the addition of indicative summaries, multimedia, and visual aids to the converted documents output by FIRST.

1.2. Readability Indices

Independent of the specific requirements of readers with ASD, readability indices are one means by which the reading difficulty of a document can be estimated. DuBay (2004) notes that over 200 readability formulae have been developed so far, with over 1 000 studies of their application published. In the research described in the present paper, the Flesch Reading Ease score (Flesch, 1949), the Kincaid readability formula (Kincaid et al., 1986), the Fog Index (Gunning, 1952), and SMOG grading (McLaughlin, 1969) metrics were selected for this purpose. Considering each in turn:
The Flesch Reading Ease score is obtained by the formula:

$$Score = 206.835 - (1.015 \times ASL) - (84.6 \times ASW)$$

Here, ASL denotes the average sentence length and ASW the average number of syllables per word. The Flesch Reading Ease Formula returns a number from 1 to 100, rather than grade level. Documents with a Flesch Reading Ease score of 30 are considered "very difficult" while those with a score of 70 are considered "easy" to read. The software developed in FIRST is therefore required to convert documents into a form with a Reading Ease Score higher than 90, commensurate with fifth grade reading level.
The Flesch-Kincaid readability formula[2] is a simplified version of the Flesch Reading Ease score. It is based on identification of the average sentence length of the document to be assessed (ASL) and the average number of syllables per word in the document (ASW). The formula estimates readability by US grade level (GL):

$$GL = (0.4 \times ASL) + (12 \times ASW) - 15$$

The Fog Index (Gunning, 1952) exploits two variables: average sentence length and the number of words containing more than two syllables ("$hard\ words$") for each 100 words of a document. This index returns the US Grade Level (GL) of the input document, according to the formula:

$$GL = 0.4 \times (average\ sentence\ length + hard\ words).$$

The SMOG grading (McLaughlin, 1969) is computed by considering the polysyllable count, equivalent to the number of words that contain more than two syllables in 30 sentences, and applying the following formula:

$$SMOG\ grading = 3 + \sqrt{polysyllable\ count}$$

It has been noted that the SMOG formula is quite widely used, particularly in the preparation of US healthcare documents intended for the general public.[3]
The selection of these standard readability metrics was made due to the observation that, although based on different types of information, they all demonstrate significant correlation in their prediction of the relative difficulty of the collections of documents assessed in the research described in this paper.
The standard readability metrics were computed using the GNU *style* package, which exploits an automatic method for syllable identification. Manual studies of the efficacy of this module suggest that it performs with an accuracy of roughly 90%, similar to state of the art part-of-speech taggers.

2. Related Work

Previous research has shown that the average US citizen reads at the seventh grade level (NCES, 1993). Experts in health literacy have recommended that materials to be read by the general population should be written at fifth or sixth grade level (Doak et al., 1996; Weiss and Coyne, 1997). The FIRST project aims to produce documents suitable for users with reading comprehension problems. Due to the reading difficulties of people with ASD, documents output by the software developed in the project should not exceed the fifth grade level (suitable for people with no reading comprehension difficulties at ten or eleven years old). Together, these constraints emphasise the desirability of consistent and reliable methods to quantify the readability of documents.
In Flesch (1949), it was found that documents presenting fictional stories lay in the range $70 \leq Score \leq 90$. Only comics were assigned a higher score for reading ease than this. The most difficult type of document was that of scientific literature, with $0 \leq Score \leq 30$. During the 1940s, the Reading Ease Scores of news articles were at the sixteenth grade level. It is estimated that in contemporary times, this has been reduced to eleventh grade level.
The set of linguistic features employed in the research described in this paper (Section 3.2.) shares some similarity with the variables shown by Gray and Leary (1935) to be

[2] To avoid confusion, in the current paper, the *Flesch-Kincaid readability formula* will hereafter be referred to as the *Kincaid readability formula*.

[3] For example, the Harvard School of Public Health provides guidance to its staff on the preparation of documents for access by senior citizens that is based on the SMOG formula (http://www.hsph.harvard.edu/healthliteracy/files/howtosmog.pdf, last accessed 1st March 2012).

closely correlated with reading difficulty. These variables include the number of first, second, and third person pronouns (correlation of 0.48), the number of simple sentences within the document (0.39), and the number of prepositional phrases occurring in the document (0.35). There is also some similarity with features exploited by Coleman (1965) in several readability formulae. These features include counts of the numbers of pronouns and prepositions occurring in each 100 words of an input document.

DuBay (2004) presents the arguments of several researchers who criticise the standard readability indices on numerous grounds. For example, the metrics have been noted to disagree in their assessment of documents (Kern, 2004). However, DuBay defends their use, arguing that the important issue is the degree of consistency that each formula offers in its predictions of the difficulty of a range of texts and the closeness with which the formulae are correlated with reading comprehension test results.

Research by Coleman (1971) and Bormuth (Bormuth, 1966) highlighted a close correlation between standard readability metrics and the variables shown to be indicative of reading difficulty. These findings motivate the current investigation into potential correlation between standard readability metrics and the metrics sensitive to the occurrence of linguistic phenomena.

3. Methodology

This section describes the methodology employed in order to explore potential correlations between the standard readability indices and the linguistic features used to measure the accessibility of different types of document for readers with ASD. It contains a description of the corpora (Section 3.1.), details of the linguistic features of accessibility that are pertinent for these readers (Section 3.2.), and details of the means by which the values of these features were automatically obtained (Section 3.3.).

3.1. Corpora

The LT developed in the FIRST project is intended to convert Bulgarian, English, and Spanish documents from fiction, news, and health genres[4] into a form facilitating the reading comprehension of users with ASD. The current paper focuses on the processing of documents written in English.

Collections of documents from these genres were compiled on the recommendation of clinical experts within the project consortium. This recommendation was based on the prediction that access to documents of these types would both motivate research into the removal of a broad spectrum of obstacles to reading comprehension and also serve to improve perceptions of inclusion on the part of readers with ASD. In the current paper, the assessment of readability is made with respect to the following document collections (Table 1):

1. **NEWS** - a collection comprising reports on court cases in the METER corpus (Gaizauskas et al., 2001) and articles from the PRESS category of the FLOB corpus.[5] The documents selected from FLOB were each of approximately 2 000 words in length. The news articles from the METER corpus were rather short; none of them had more than 1 000 words. We included only documents with at least 500 words;
2. **HEALTH** - a collection comprising healthcare information contained in a collection of leaflets for distribution to the general public, from categories *A01*, *A0J*, *B1M*, *BN7*, *CJ9*, and *EDB* of the British National Corpus (Burnard, 1995). This sample contains documents with considerable variation in word length;
3. **FICTION** - a collection of documents from the FICTION category of the FLOB corpus. Each is approximately 2 000 words in size; and
4. **SIMPLEWIKI** - a random selection of simplified **encyclopaedic** documents, each consisting of more than 1 000 words, from Simple Wikipedia.[6] This collection is included as a potential model of accessibility. One of the goals of the research described in this paper is to compare the readability of other types of document from this "standard".

Corpus	Words	Texts
SimpleWiki	272,445	170
News	299,685	171
Health	113,269	91
Fiction	243,655	120

Table 1: Size of the corpora

3.2. Linguistic Features of Document Accessibility

The obstacles to reading comprehension faced by people with ASD when seeking to access written information were presented in Section 1.1. The features presented in this section are intended to indicate the occurrence of these obstacles in input documents. Thirteen features are proposed as a means of detecting the occurrence of the different types of obstacle to reading comprehension listed in Section 1.1. Related groups of features are presented below.

(1) **Features indicative of structural complexity**: This group of ten features was inspired by the syntactic concept of the projection principle (Chomsky, 1986) that "lexical structure must be represented categorically at every syntactic level". This implies that the number of noun phrases in a sentence is proportional to the number of nouns in that sentence, the number of verbs in a sentence is related to the number of clauses and verb phrases, etc. The values of nine of these features were obtained by processing the output of *Machinese Syntax*[7] to detect the

[4] In this paper, we use the term *health* to denote documents from the genre of education in the domain of health.

[5] Freiburg-LOB Corpus of British English (http://khnt.hit.uib.no/icame/manuals/flob/INDEX.HTM)

[6] http://simple.wikipedia.org

[7] http://www.connexor.eu

Feature	Indicator of
Nouns (N)	References to concepts/entities
Adjectives (A)	Descriptive information about concepts/entities
Determiners (Det)	References to concepts that are not proper names, acronyms, or abbreviations
Adverbs (Adv)	Descriptive information associated with properties of and relations between concepts/entities
Verbs (V)	Properties of and relations between concepts/entities
Infinitive markers (INF)	Infinitive verbs (a measure of syntactic complexity)
Coordinating conjunctions (CC)	Coordinated phrases
Subordinating conjunctions (CS)	Subordinated phrases, including phrases embedded at multiple levels
Prepositions (Prep)	Prepositional phrases (a well-cited source of syntactic ambiguity and complexity)

Table 2: Features (structural complexity)

occurrence of words/lemmas with particular part-of-speech tags (Table 2). As the tenth feature we proposed *Sentence complexity* (Compl) in terms of number of verb chains. It was measured as the ratio of the number of sentences in the document containing at most one verb chain to the number containing two or more verb chains. To illustrate, the sentence:

> *I am consumed with curiosity, and I cannot rest until I know why this Major Henlow should have sent the Runners after you.*

contains four verb chains: {*am consumed*}, {*cannot rest*}, {*know*}, and {*should have sent*}. This feature exploits the functional tags assigned to different words by *Machinese Syntax* (Section 3.3.1.).

(2) **Features indicative of ambiguity in meaning**: This group of three features (Table 3) is intended to indicate the amount of semantic ambiguity in the input document.

Feature	Indicator of
Pronouns (Pron)	Anaphoric references
Definite descriptions (defNP)	Anaphoric references
Word senses (Senses)	Semantic ambiguity

Table 3: Features (ambiguity in meaning)

In all three cases, the difficulties caused by the feature arise as a result of doubts over the reference to concepts in the domain of discourse by different linguistic units (words and phrases). The values of these features are obtained by processing the output of *Machinese Syntax* to detect both the occurrence of words/lemmas with particular parts of speech and the functional dependencies holding between different words, and exploitation of WordNet as a source of information about the senses associated with content words in the input text.
These features were calculated as averages per sentence. The only exception was the feature *Senses* which was computed as the average number of senses per word.

3.3. Extraction of Linguistic Features

A user requirements analysis undertaken during the initial stage of the project motivated the development of features of accessibility based on the occurrence of various linguistic phenomena in an input document. Given that these are complex and difficult to detect automatically, the linguistic features are based on indicative morpho-syntactic information that can be obtained via existing NLP resources.
Derivation of the feature values depends on exploitation of two language technologies: Connexor's *Machinese Syntax* functional dependency parser (Tapanainen and Jarvinen, 1997) and the generic ontology, WordNet (Fellbaum, 1998). The detection process is based on the assumption that words with particular *morphological* and *surface syntactic* tags assigned by *Machinese Syntax* indicate the occurrence of different types of linguistic phenomenon.
One caveat that should be made with regard to the values obtained for these features is that they exploit language processing technology that is imperfect in its accuracy and coverage. The efficacy of Connexor's *Machinese Syntax*, used to obtain the values for the linguistic features, is described in (Tapanainen and Jarvinen, 1997).

3.3.1. Functional Dependencies

The values of two features, *defNP* and *Compl*, are obtained by reference to the functional dependencies detected by *Machinese Syntax* between words in the input documents.
The feature defNP is intended to obtain the number of definite noun phrases occurring in each sentence of an input document. This number is measured by counting the number of times that functional dependencies occur between tokens with the lemma *the*, *this*, and *that* and tokens with a nominal surface syntactic category.
The feature *Compl*, which relies on identification of the verb chains occurring in each sentence of a document (see Section 3.2.), exploits analyses provided by the parsing software. Verb chains are recognised as cases in which verbs are assigned either *finite main predicator* or *finite auxiliary predicator* functional tags by *Machinese Syntax*.

3.3.2. WordNet

Word sense ambiguity (*Senses*) was detected by exploitation of the WordNet ontology (Fellbaum, 1998). Input documents are first tokenised and each token disambiguated in terms of its surface syntactic category by *Machinese Syntax*. The number of concepts linked to the word when used with that category were then obtained from WordNet. The extraction method thus exploits some limited word sense disambiguation as a result of the operation of the parser. As noted earlier (Section 3.2.), the feature *Senses* was calculated as the average number

Corpus	Kincaid	Flesch	Fog	SMOG	ch/w	syl/w	w/s
SimpleWiki	**7.49**	**69.91**	**10.35**	**9.78**	**4.67**	**1.43**	**16.05**
News	**9.39**	**64.98**	**12.28**	**10.77**	4.66	1.43	**20.90**
Health	7.84	69.31	10.83	10.07	4.63	1.42	**17.13***
Fiction	**5.05**	**83.06**	**7.85**	**7.90**	**4.29**	**1.30**	**13.58**

Table 4: Readability indices and related features

of senses per word. Therefore, multiple occurrences of the same ambiguous word will increase the value of this feature.

4. Results

The study presented in this paper comprises three parts. In the first, a comparison is made between the values obtained for the four readability indices and the factors that they exploit (average numbers of characters and syllables per word, average number of words per sentence) in their assessment of the corpora (SimpleWiki, News, Health, and Fiction). If the intuitive assumption is valid, that SimpleWiki represents a corpus of simplified texts (a "gold standard"), then this comparison will indicate how far documents from the news, health, and fiction genres (important for the social inclusion of people with ASD) lie from this 'gold standard'.

In the second part, the use of thirteen linguistic features is explored. Ten of the linguistic features are based on the frequency of occurrence of surface syntactic tags, one is based on sentence complexity expressed in terms of the number of verb chains that they contain, another provides an approximation of the number of definite noun phrases used in the text, and the final feature measures the average level of semantic ambiguity of the words used in the text. The values obtained for these features for each of the corpora are compared.

In the third part of the study, potential correlations between the linguistic features and readability metrics are investigated. The motivation for this lies in the fact that extraction of the linguistic features is relatively expensive and unreliable, while the computation of the readability metrics is done automatically and with greater accuracy. The ability to estimate the accessibility of documents for people with ASD on the basis of easily computed readability metrics rather than complex linguistic features would be of considerable benefit.

The results obtained in these three parts of the study are presented separately in the following sections.

4.1. Readability

The results of the first part of this study are presented in Table 4. The first row of the table contains the scores for these seven features obtained for SimpleWiki. For the other three text genres, the values of these features were calculated and a non-parametric statistical test (Kolmogorov-Smirnov Z test) was applied in order to calculate the significance of the differences in means between SimpleWiki and the corresponding text genre.[8]

In Table 4, values which differ from those obtained for the documents in SimpleWiki at a 0.01 level of significance are printed in bold. Those printed in bold with an asterisk differ from those obtained from documents in SimpleWiki at a 0.05, but not at a 0.01 level of significance.

On the basis of these results it can be inferred that the news texts are most difficult to read as they require a higher level of literacy for their comprehension (the values of the Kincaid, Fog and SMOG indices are maximal for this genre, while the Flesch index is at its lowest level, indicating that all are in accordance). Discrepancies between the values of different indices are not surprising, as they use different variables and different criterion scores (DuBay, 2004). Also, it is known that the predictions made by these formulae are not perfect, but are rough estimates (r = .50 to .84) of text difficulty. That is, they "account for 50 to 84 percent fo the variance in text difficulty as measured by comprehension tests" (DuBay, 2004). In the context of the current research, it is important that when the difficulty of two types of text is compared, consistent conclusions can be made about which type is more difficult than the other, regardless of which readability formula is used (Table 4).

It is interesting to note that none of the indices indicate significant differences between the readability of health documents and that of documents in SimpleWiki, suggesting that similar levels of literacy are necessary for their comprehension. Despite this, a slightly greater average sentence length was noted for the health texts than the texts from SimpleWiki. The most surprising finding was that the fiction texts are reported by all readability metrics (including average word and sentence length) to be significantly less difficult than those from SimpleWiki (Table 4). These results cast doubt on the assumption that SimpleWiki serves as a paradigm of accessibility, which has been made in previous work on text simplification (e.g. (Coster and Kauchak, 2011)).

As it was observed that all readability indices returned similar values in the comparison of different text genres (Table 4), the strength of correlation between them was investigated. To this end, Pearson's correlation was calculated between each pair of the four indices (Table 6), over the whole corpora (SimpleWiki, News, Health, and Fiction). Pearson's correlation is a bivariate measure of strength of the relationship between two variables, which can vary from 0 (for a random relationship) to 1 (for a perfectly linear relationship) or -1 (for a perfectly negative linear relationship). The results presented in Table 6 indicate a very strong linear correlation between each pair of readability indices.

[8]The Kolmogorov-Smirnov Z test was selected as a result of prior application of the Shapiro-Wilk's W test which demonstrated that most of the features do not follow a normal distribution.

Corpus	V	N	Prep	Det	Adv	Pron	A	CS	CC	INF	Compl	Senses	defNP
SimpleWiki	**2.74**	**5.77**	**1.92**	**1.94**	**0.77**	**0.81**	**1.20**	**0.19**	**0.60**	**0.21**	**1.37**	**6.59**	**1.19**
News	**4.08**	**6.63**	**2.44**	**2.17**	**0.95**	**1.64**	**1.56**	**0.35**	**0.65***	**0.38**	**0.53**	**6.73***	**1.26***
Health	**3.40**	**5.22**	1.82	**1.51**	**0.99**	**1.38**	**1.63**	**0.32**	**1.01**	**0.35**	**1.15**	6.73	**0.73**
Fiction	2.95	**3.33**	**1.43**	**1.35**	**1.10**	**1.89**	**0.90**	**0.23**	**0.49**	**0.23***	**1.13***	**7.59**	**0.77**

Table 5: Linguistic features

r	Kincaid	Flesch	Fog	SMOG
Kincaid	1	−.959	.987	.951
Flesch	−.959	1	−.957	−.972
Fog	.987	−.957	1	.979
SMOG	.951	−.972	.979	1

Table 6: Pearson's correlation between readability indices

In the case of the Flesch index, the higher the score is, the lower the grade level necessary for understanding the given text. For all other indices, a higher score indicates a higher grade level necessary to understand the text. Therefore, the correlation between the Flesch index and any other index is always reported as negative. In order to confirm that these correlations are not genre dependent, a set of experiments was conducted to measure the correlation between these four readability measures separately for each of the four corpora (SimpleWiki, News, Health and Fiction). Those experiments revealed a very close correlation between the four readability indices (between .915 and .993) in each genre.

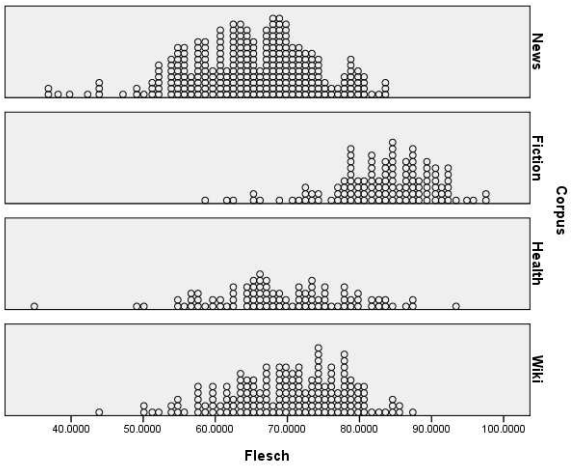

Figure 1: Distribution of the Flesch index

As the correlation between the four readability indices was reported to be almost perfectly linear (Table 6), the reminder of this study focuses on the Flesch index as a representative of the readability indices. The results discussed earlier (Table 4) presented only the mean value of the Flesch index for each of the corpora. In Figure 1, each text is represented separately, providing a more complete picture of the Flesch index distribution across the corpora.

It can be noted that the mean value of the Flesch index lies at approximately the same place on the x-axis for both SimpleWiki and Health texts, which is in accordance with the previously reported results (Table 4). Mean value of the Flesch index in News genre slightly shifted to the left relative to the SimpleWiki corresponds to lower text readability in News genre than in SimpleWiki reported in Table 4. It can also be noted that the distribution of the Flesch index in the Fiction genre is positioned significantly to the right relative to the SimpleWiki, thus indicating a higher readability of texts in this genre.

4.2. Linguistic Features

The investigation of the average occurrence of ten different POS tags per sentence (V[9], N, Prep, Det, Adv, Pron, A, CS, CC, INF) and three other linguistically motivated features (Compl, Senses and defNP) showed significantly different values in News, Health and Fiction than in SimpleWiki in most of the cases (Table 5).

Documents from SimpleWiki were found to contain the *highest* ratio between simple and complex sentences (Compl), the *lowest* number of verbs (V), adverbs (Adv), pronouns (Pron), subordinating conjunctions (CS), infinitive markers (INF) and senses per word (Senses), which may reflect a certain simplicity of these texts.

The News genre was found to contain the *lowest* ratio of simple to complex sentences (Compl), and the *highest* number of verbs (V), subordinate conjunctions (CS) and infinitive markers (INF) per sentence. These features indicate a greater number of verb chains (Compl) and subordinate clauses (CS), longer verb chains and more complex verb constructions (V and INF) for news articles. These features can be considered indicators of syntactic complexity, which is probably reflected in the high scores for readability indices obtained in this genre (Table 4). The texts from the genre of fiction contained the smallest average number of nouns (N), prepositions (Prep), determiners (Det), adjectives (A) and coordinating conjunctions (CC) per sentence (Table 5). However, this genre contained a significantly higher number of senses per word (Senses) than other genres.

4.3. Flesch vs. Linguistic Features

In the third part of the study, potential correlation between the linguistic features and readability indices was investigated. The Flesch index was selected as a representative of readability indices (as all four readability indices were almost perfectly linearly correlated, selection of an alternative readability index should not change the results significantly). Pearson's correlation between the investigated POS frequencies (on average per sentence)

[9]This tag includes the occurrence of present (ING) and past participle (EN).

Corpus	V	N	Prep	Det	Adv	Pron	A	CS	CC	INF
all	**−.493**	**−.812**	**−.777**	**−.715**	**−.093***	**.189**	**−.769**	−.377	**−.464**	**−.415**
SimpleWiki	**−.397**	**−.552**	**−.641**	**−.545**	**−.293**	.136	**−.685**	−.130	**−.424**	−.118
News	**−.385**	**−.738**	**−.759**	**−.705**	**−.197**	**.291**	**−.783**	**−.438**	**−.387**	**−.426**
Health	**−.274**	**−.743**	**−.607**	**−.489**	−.104	.078	**−.703**	.014	**−.610**	−.139
Fiction	**−.605**	**−.889**	**−.854**	**−.851**	**−.555**	−.146	**−.876**	**−.515**	**−.670**	**−.506**

Table 7: Pearson's correlation between Flesch readability index and POS frequencies

Corpus	Compl	ch/w	syl/w	w/s	Senses	defNP
All	**.210**	**−.859**	**−.922**	**−.792**	**.627**	**−.595**
SimpleWiki	**.209**	**−.825**	**−.921**	**−.643**	**.452**	**−.337**
News	−.026	**−.866**	**−.919**	**−.762**	**.568**	**−.688**
Health	0.034	**−.771**	**−.918**	**−.705**	**.417**	**−.450**
Fiction	**.376**	**−.790**	**−.877**	**−.822**	**.738**	**−.791**

Table 8: Pearson's correlation between Flesch readability index and other features

and the Flesch index is presented in Table 7, while the correlation between the other six features and the Flesch index is reported in Table 8. These experiments were conducted first for all the corpora and then for each corpus separately in order to determine whether these correlations may be genre dependent.

As would be expected, the direction of correlation (sign '−' or '+') is independent of genre (in those cases where the correlation is statistically significant and thus more reliable). However, the strength of the correlation does depend of the genre of the texts, e.g. correlation between average number of verbs per sentence (V) and the Flesch index is −.274 for the Health and −.605 for the Fiction genres. The '−' sign indicates that if the value of the feature increases, the Flesch index decreases (indicating a less readable text) and vice-versa (as the Pearson's correlation is a symmetric function we are not able to say in which direction the correlation goes). The results presented in Tables 7 and 8 therefore indicate that for most of the features (V, N, Prep, Det, Adv, A, CS, CC, INF, ch/w, w/s, defNP) the lower the feature value for a given text, the easier that text is to read (the higher the Flesch index). For feature Compl, the results also support the intuition that the higher the ratio of simple to complex sentences is in the text, the more readable it is (higher Flesch index).

The most surprising results were those obtained for the feature Senses (Table 8), which indicate that the higher the average number of senses per word in the text, the more readable the text is. One possible hypothesis that emerges from this observation is that shorter words in English tend to be more semantically ambiguous than longer words (the readability indices are highly correlated with word length, measured both in characters and syllables per word, with the occurrence of shorter words suggesting that the text is easier to read).

5. Conclusions

There are several important findings of this study. First, it was shown that the four well-known readability indices are almost perfectly linearly correlated on each of the four investigated text genres – SimpleWiki, News, Health, and Fiction. Furthermore, our results indicated that texts from the genre of fiction are simpler than those selected from SimpleWiki in terms of the readability indices, casting doubt on the assumption that SimpleWiki is a useful source of documents to form a gold standard of accessibility for people with reading difficulties. Application of the measures also indicated that news articles are most difficult to read, relative to the other genres, requiring a higher level of literacy for their comprehension.

The results of the second part of our study (investigation of various linguistic features) revealed that documents from SimpleWiki were the simplest of the four corpora in terms of several linguistic features – average number of verbs, adverbs, pronouns, subordinate conjunctions, infinitive markers, number of different word senses and ratio between simple and complex sentences. They also indicated some of the factors that may make news texts difficult to read, e.g. containing the highest numbers of verbs and subordinate conjunctions per sentence, and the lowest ratio of simple to complex sentences.

The results of the third set of experiments indicated the average length of words (in characters and in syllables) as being features with the highest correlation to the Flesch index. They also indicated that features such as the average number of nouns, prepositions, determiners and adjectives are closely correlated with the Flesch index (up to .89 in the fiction genre), which supports the idea of using readability indices as an initial measure of text complexity in our project. The comparison of these correlations across different text genres demonstrated that they are genre dependent and that the correlation between these linguistic features and the Flesch index is closest for the Fiction genre.

6. Acknowledgements

The research described in this paper was partially funded by the European Commission under the Seventh (FP7 - 2007-2013) Framework Programme for Research and Technological Development (FIRST 287607). This publication [communication] reflects the views only of the authors, and the Commission cannot be held responsible for

any use which may be made of the information contained therein.

7. References

J. R. Bormuth. 1966. Readability: A new approach. *Reading research quarterly*, 1:79–132.

W. Bosma, 2005. *Image retrieval supports multimedia authoring*, pages 89–94. ITC-irst, Trento, Italy.

L. Burnard. 1995. *Users Reference Guide British National Corpus Version 1.0*. Oxford University Computing Services, UK.

N. Chomsky. 1986. *Knowledge of language: its nature, origin, and use*. Greenwood Publishing Group, Santa Barbara, California.

E. B. Coleman, 1965. *On understanding prose: some determiners of its complexity*. National Science Foundation, Washington, D.C.

E. B. Coleman, 1971. *Developing a technology of written instruction: some determiners of the complexity of prose*. Teachers College Press, Columbia University, New York.

W. Coster and D. Kauchak. 2011. Simple english wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011)*, pages 665–669, Portland, Oregon, June. Association of Computational Linguistics.

C. C. Doak, L. G. Doak, and J. H. Root. 1996. *Teaching patients with low literacy skills*. J. B. Lippincott Company, Philadelphia.

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

G. Escudero, L. Màrquez, and G. Rigau. 2000. A comparison between supervised learning algorithms for word sense disambiguation. In C. Cardie, W. Daelemans, C. Nédellec, and E. Tjong Kim Sang, editors, *Proceedings of the Fourth Computational Natural Language Learning Workship, CoNLL-2000*, pages 31–36, Lisbon, Portugal, September. Association of Computational Linguistics.

R. Evans. 2011. Comparing methods for the syntactic simplification of sentences in information extraction. *Literary and Linguistic Computing*, 26 (4):371–388.

C. Fellbaum. 1998. *WordNet An Electronic Lexical Database*. MIT Press, Cambridge, MA.

R. Flesch. 1949. *The art of readable writing*. Harper, New York.

U. Frith and M. Snowling. 1983. Reading for meaning and reading for sound in autistic and dyslexic children. *Journal of Developmental Psychology*, 1:329–342.

R. Gaizauskas, J. Foster, Y. Wilks, J. Arundel, P. Clough, and S. Piao. 2001. The Meter corpus: A corpus for analysing journalistic text reuse. In *Proceedings of Corpus Linguistics 2001 Conference*, pages 214–223. Lancaster University Centre for Computer Corpus Research on Language.

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

R. Gunning. 1952. *The technique of clear writing*. McGraw-Hill, New York.

R. P. Kern. 2004. *Usefulness of readability formulas for achieving Army readability objectives: Research and state-of-the-art applied to the Army's problem (NTIS No. AD A086 408/2)*. U.S. Army Research Institute, Fort Benjamin Harrison.

J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. 1986. *Derivation of new readability formulas (Automatic Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel*. CNTECHTRA.

G. H. McLaughlin. 1969. SMOG grading – a new readability formula. *Journal of reading*, 22:639–646.

N. Minshew and G. Goldstein. 1998. Autism as a disorder of complex information processing. *Mental Retardation and Developmental Disability Research Review*, 4:129–136.

R. Mitkov. 2002. *Anaphora Resolution*. Longman, Harlow, Essex.

K. Nation, P. Clarke, B. Wright, and C. Williams. 2006. Patterns of reading ability in children with autism-spectrum disorder. *Journal of Autism & Developmental Disorders*, 36:911–919.

NCES. 1993. *Adult literacy in America*. National Center for Education Statistics, U.S. Dept. of Education, Washington, D.C.

I. M. O'Connor and P. D. Klein. 2004. Exploration of strategies for facilitating the reading comprehension of high-functioning students with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 34:2:115–127.

C. Orăsan and L. Hasler. 2007. Computer-aided summarisation: how much does it really help? In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2007)*, pages 437–444, Borovets, Bulgaria, September.

A. Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4:1:77–109.

P. Tapanainen and T. Jarvinen. 1997. A non-projective dependency parser. In *Proceedings of the 5th conference on Applied Natural Language Processing of the Association for Computational Linguistics*, pages 64–71. Association of Computational Linguistics.

B. D. Weiss and C. Coyne. 1997. The use of user modelling to guide inference and learning. *New England Journal of Medicine*, 337 (4):272–274.

M. Yatskar, B. Pang, C. Danescu-Niculescu-Mizil, and L. Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from wikipedia. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pages 365–368, Los Angeles, California, June. Association of Computational Linguistics.