Towards an Improved Methodology for Automated Readability Prediction

Philip van Oosten, Dries Tanghe, Véronique Hoste

LT^3 Language and Translation Technology Team, University College Ghent
Groot-Brittanniëlaan 45, 9000 Gent, Belgium
{philip.vanoosten, dries.tanghe, veronique.hoste}@hogent.be

Abstract

Since the first half of the 20th century, readability formulas have been widely employed to automatically predict the readability of an unseen text. In this article, the formulas and the text characteristics they are composed of are evaluated in the context of large Dutch and English corpora. We describe the behaviour of the formulas and the text characteristics by means of correlation matrices and a principal component analysis, and test the methodological validity of the formulas by means of collinearity tests. Both the correlation matrices and the principal component analysis show that the formulas described in this paper strongly correspond, regardless of the language for which they were designed. Furthermore, the collinearity test reveals shortcomings in the methodology that was used to create some of the existing readability formulas. All of this leads us to conclude that a new readability prediction method is needed. We finally make suggestions to come to a cleaner methodology and present web applications that will help us collect data to compile a new gold standard for readability prediction.

1. Introduction

The concept of readability has been defined in a wide variety of ways, typically dependent on the author's intentions. For instance, Staphorsius (1994) defines readability of a text as the reading proficiency that is needed for text comprehension. The author's intention of designing a formula to determine the suitability of reading material given a certain reading proficiency is not without its influence in that definition. McLaughlin (1974), the author of the influential SMOG formula, on the other hand, defines readability as the characteristic of a text that makes readers willing to read on.

In our paper, there is no decisive definition of the concept of readability. However, we follow a strong tradition when we take as our working definition "what makes some texts easier to read than others." (DuBay, 2004)

Our goal is to find correspondences between readability scores and text characteristics when those are calculated for each text in a large corpus. Readability scores calculated with different readability formulas strongly correspond. Furthermore, the methodology that was used to construct some formulas is unsound and it is very unlikely that those formulas predict readability at all.

In the following section, we present a number of formulas that assign a holistic readability score to any text. In section 3. results of experiments show that these formulas correlate very strongly with superficial text characteristics, and even moreso with each other. Further, the validity of the methodology used to obtain the readability formulas is assessed. In section 4. we argue that a new methodology for constructing readability prediction systems is needed, and that we therefore need a new gold standard corpus. We present web applications that will help us in compiling such a gold standard.

2. Existing readability formulas

A *readability formula* is a mathematical formula intended for indicating the difficulty of text. The formula typically consists of a number of variables, which are characteristics of the text, and constant weights. We briefly explain the text characteristics that figure as variables in the formulas discussed in this article in table 1. A value for the variables can be obtained by automatically processing a text. In the rest of this article, the terms *text characteristic* and *variable* are used interchangeably.

Name	Description
avgnumsyl	Average word length in number of syllables.
avgsentencelen	Average sentence length in number of words.
avgwordlen	Average word length in number of characters.
freq77	Percentage of words also found in a Dutch word list with a cumulative frequency of 77%. The list is based on a list ordered by descending frequency in the "27 Miljoen Woorden Krantencorpus 1995", which is available through the HLT agency at `http://tst.inl.nl/en/producten`.
psw	Percentage of sentences per word.
ttr	Type/token ratio, the number of unique words divided by the total number of words.
freq3000	Percentage of words not on the Dale-Chall (1948) word list. The Dale-Chall word list contains 3000 of the most frequent words in the English language.
avgpolysylsent	Average number of words of 3 or more syllables per sentence.
ppolysylword	Percentage of words of three or more syllables.
ratiolongword	Ratio of words of more than 6 characters.

Table 1: Variables that are used in the readability formulas in table 2. All variables are derived on a text by text basis.

The readability formulas discussed in this article are listed in table 2. For some formulas, a higher score applies to a more difficult text and a lower score to a more readable text. Their slope is considered positive. For the other formulas, the situation is exactly opposite and the slope is considered negative.

For our experiments, we implemented an interface to automate the computation of a number of existing readability

Dutch

Name	Acronym	Formula	Slope
CLIB: Cito leesbaarheidsindex voor het basisonderwijs (Staphorsius, 1994)	*clib*	$46 + 0.47 \times freq77 - 6.6 \times avgwordlen$ $-0.37 \times ttr + 1.4 \times psw$	+
CILT: Cito leesindex technish lezen (Staphorsius, 1994)	*cilt*	$114 + 0.28 \times freq77 - 12 \times avgwordlen$	+
Flesch-Douma (Douma, 1960)	*douma*	$207 - 0.93 \times avgsentencelen - 77 \times avgnumsyl$	-
Leesindex Brouwer (Brouwer, 1963)	*brouwer*	$195 - 2 \times avgsentencelen - 67 \times avgnumsyl$	-

English

Name	Acronym	Formula	Slope
Flesch Reading Ease (Flesch, 1948)	*flesch*	$207 - avgsentencelen - 85 \times avgnumsyl$	-
Dale-Chall Reading Grade Score (Dale and Chall, 1948)	*rgs*	$0.16 \times freq3000 + 0.05 \times avgsentencelen + 3.6$	+
Coleman-Liau Index (Coleman and Liau, 1975)	*cli*	$5.9 \times avgwordlen - 0.3 \times avgsentencelen - 16$	+
Flesch-Kincaid Grade Level (Kincaid et al., 1975)	*kincaid*	$0.39 \times avgsentencelen + 12 \times avgnumsyl - 16$	+
Gunning Fog Index (Gunning, 1952)	*fog*	$0.4 \times (avgsentencelen + ppolysylword)$	+
ARI: Automated Readability Index (Senter and Smith, 1967)	*ari*	$4.7 \times avgwordlen + 0.5 \times avgsentencelen - 21$	+
SMOG: Simple Measure of Gobbledygook (McLaughlin, 1969)	*smog*	$\sqrt{30 \times avgpolysylsent} + 3.1$	+

Swedish

Name	Acronym	Formula	Slope
Läsbarhetsindex Björnsson (Björnsson, 1968)	*lix*	$avgsentencelen + ratiolongword$	+

Table 2: Readability formulas. The constant factors are rounded to 2 significant digits. The meaning of the variables used in the readability formulas is explained in table 1. The slope is indicated with a + for formulas that indicate a greater difficulty by means of a higher score and with a - otherwise.

formulas.[1] The formulas discussed here are designed for either Dutch (Douma, 1960; Brouwer, 1963; Staphorsius, 1994), English (Dale and Chall, 1948; Flesch, 1948; Gunning, 1952; Senter and Smith, 1967; McLaughlin, 1969; Coleman and Liau, 1975; Kincaid et al., 1975) or Swedish (Björnsson, 1968), as indicated in table 2.

In order to count syllables, which is needed for *avgnumsyl*, *ppolysylword* and *avgpolysylsent*, a classification based syllabifier was implemented. The procedure followed, which amounts to training an automatic syllabification algorithm on pronunciation dictionaries, was first put forward by Daelemans and van den Bosch (1992). In our implementation, the dictionaries used were taken from the CELEX-database (Baayen et al., 1995). The number of syllables is correct for 96.5 percent of all words, both for Dutch and English. That accuracy was established using 10-fold cross validation on CELEX.

3. Experiments

3.1. Data sets

In order to investigate the correspondences between the readability formulas and variables in table 2, we used four different corpora in two different languages: Dutch and English.

3.1.1. Dutch corpora

Eindhoven Corpus (ehc, 1989) The Eindhoven Corpus is a corpus that aims to present a cross section of Dutch language usage. In nearly 5000 text fragments, it contains a total of 740k tokens, divided over the following six text types: newspapers, opinion, family and popular science magazines, literary works, and official documents. The Eindhoven Corpus also contains a section of transcribed speech, which we disregarded in our experiments on the grounds that our research is primarily focused on the readability of written text. The acronym EHC will further refer to the Eindhoven Corpus.

SoNaR (Schuurman et al., 2009) SoNaR is a general Dutch reference corpus that is still under development. The final version will contain at least 500 million tokens, which will encompass as many aspects of written Dutch as possible. We have worked with a preliminary release of almost 81 million tokens in approximately 213k texts. Titles and headers were not counted, and we disregarded texts that were written to be spoken or that are unstructured, such as autocues and discussion lists.

3.1.2. English corpora

British National Corpus (British National Corpus Consortium, 2000) The British National Corpus is composed of British English text samples of greatly varying length. It includes both imaginative and informative texts and is not

[1] A documented demo of our implementation is available at `http://lt3.hogent.be/demos/readability`.

specifically restricted to any particular subject field, register or genre. It contains examples of both spoken and written language, but in this article, only the written texts are taken into account. Over 3100 texts from the British National Corpus are incorporated in the data set, containing a total of almost 85 million tokens, not counting titles and headings. The acronym BNC will be used to refer to this corpus.

Penn Treebank (Marcus et al., 1993) The Penn Treebank consists of articles from the Wall Street Journal and a collection of unpublished written communication. We used the Wall Street Journal section of the Penn Treebank, which contains over 1 million tokens in 2500 texts. We further refer to that section by the acronym WSJ.

3.2. Experimental setup

For each processed text, the value of the characteristics that are used in the readability formulas and their results were determined. In order to compute the correspondences between different readability formulas and text characteristics for each corpus, a correlation metric is needed. The Pearson correlation coefficient (StatSoft Inc., 2007) is a standard metric to indicate the correlation between two variables. However, one of the assumptions it is based on is that those variables are normally distributed. We used Shapiro-Wilks' W-test (Royston, 1982), which is a standard test to determine whether a statistical variable is normally distributed. That test indicated that generally neither readability scores nor text characteristics are normally distributed along the texts of a corpus. Therefore, Pearson is inappropriate and we opted for a non-parametric statistic, namely the Spearman rank correlation.

To further substantiate the analysis of the correlations, we carried out a *principal component analysis* (PCA) (Baayen, 2008) on both the readability scores and the text characteristics. PCA attempts to determine how many independent factors cause the distribution in a data set. By means of PCA on the readability scores, we tried to deduce how many independent factors account for the results of the different readability formulas. That should give an idea to what degree different formulas are independent. Similarly, through PCA on the text characteristics, we attempted to establish how many characteristics can be regarded as independent. Since *freq3000* is not meaningful for Dutch, it was not considered for the PCA of the text characteristics for the Dutch corpora and because *freq77* is meaningless for English texts, it was not taken into account for the English corpora.

In order to certify whether the design of the readability formulas is still valid for the corpora employed in this article, a *collinearity* test as described by Belsley et al. (1980) was performed for all formulas, except *smog*[2]. Such tests determine whether variables used in a multiple regression formula linearly depend on each other. If that is the case, the result of the multiple regression analysis can probably not be extrapolated. In case of readability formulas which are the result of a multiple regression, this means that the formulas may not be applicable to other texts than those used in the corpus used to construct them. In short, such formulas can not be used to predict readability.

Key	Formula or text characteristic
1	*brouwer*
2	*kincaid*
3	*fog*
4	*smog*
5	*ari*
6	*douma*
7	*flesch*
8	*lix*
9	*cli*
10	*cilt*
11	*clib*
12	*rgs*
13	*avgpolysylsent*
14	*avgwordlen*
15	*avgnumsyl*
16	*ppolysylword*
17	*ratiolongword*
18	*psw*
19	*avgsentencelen*
20	*freq77* (Dutch corpora) or *freq3000* (English corpora)
21	*ttr*

Table 3: Key to the numbers in the headers of table 4.

3.3. Experimental results

Table 4 shows the Spearman correlation matrices for the four corpora. The key to the numbers in the header row and the header column can be found in table 3. The correlation between the *flesch* and *kincaid* formulas, for example, is displayed in the crossing of row 7 and column 2 for each corpus. The corpora are identified as follows: part A refers to EHC, part B to SoNaR, part C to BNC and part D to WSJ.

A correlation in itself does not express a causal relationship between two different properties. It is therefore necessary to explain the correlations in further detail.

In the three following paragraphs, we will discuss the correlations between text characteristics (3.3.1.), correlations between readability formulas and the text characteristics used therein (3.3.2.) and the correlations between readability formulas (3.3.3.).

3.3.1. Correlations between text characteristics

Word length and word frequency Strong correlations are observed between text characteristics related to word length, both when measured in number of syllables or in number of characters. Those text characteristics are *avgnumsyl* (15 in table 4), *ppolysylword* (16), *avgwordlen* (14) and *ratiolongword* (17). In BNC (part C in table 4) and EHC (part A), such high correlations also extend to *avgpolysylsent* (13), which depends on both word length and the number of sentences per text.

Generally, variables that measure word length strongly correlate with each other, because their values originate from the same property of text. Naturally, if the word length

[2] The collinearity test can not be performed on *smog* because only one variable occurs in that formula.

A	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1		-1.00	-0.98	-0.97	-0.99	0.96	0.96	-0.96	-0.85	0.74	0.76	-0.62	-0.97	-0.77	-0.81	-0.77	-0.73	0.83	-0.83	0.27	-0.13
2	-1.00		0.98	0.97	0.99	-0.95	-0.95	0.96	0.84	-0.72	-0.76	0.63	0.97	0.75	0.79	0.76	0.72	-0.85	0.85	-0.26	0.12
3	-0.98	0.98		1.00	0.96	-0.93	-0.93	0.96	0.82	-0.71	-0.74	0.61	1.00	0.74	0.78	0.81	0.74	-0.83	0.83	-0.27	0.11
4	-0.97	0.97	1.00		0.96	-0.93	-0.93	0.96	0.82	-0.71	-0.74	0.61	1.00	0.73	0.77	0.80	0.73	-0.84	0.84	-0.26	0.11
5	-0.99	0.99	0.96	0.96		-0.92	-0.92	0.96	0.84	-0.73	-0.77	0.64	0.96	0.75	0.75	0.72	0.71	-0.87	0.87	-0.28	0.13
6	0.96	-0.95	-0.93	-0.93	-0.92		1.00	-0.91	-0.94	0.86	0.81	-0.56	-0.93	-0.89	-0.94	-0.89	-0.84	0.66	-0.66	0.32	-0.19
7	0.96	-0.95	-0.93	-0.93	-0.92	1.00		-0.91	-0.94	0.86	0.81	-0.56	-0.93	-0.89	-0.94	-0.89	-0.84	0.66	-0.66	0.32	-0.19
8	-0.96	0.96	0.96	0.96	0.96	-0.91	-0.91		0.83	-0.74	-0.77	0.62	0.96	0.75	0.76	0.76	0.81	-0.82	0.82	-0.31	0.14
9	-0.85	0.84	0.82	0.82	0.84	-0.94	-0.94	0.83		-0.96	-0.86	0.48	0.82	0.99	0.95	0.90	0.90	-0.50	0.50	-0.39	0.28
10	0.74	-0.72	-0.71	-0.71	-0.73	0.86	0.86	-0.74	-0.96		0.89	-0.39	-0.71	-0.98	-0.93	-0.88	-0.90	0.35	-0.35	0.58	-0.31
11	0.76	-0.76	-0.74	-0.74	-0.77	0.81	0.81	-0.77	-0.86	0.89		-0.46	-0.74	-0.84	-0.79	-0.75	-0.79	0.51	-0.51	0.63	-0.53
12	-0.62	0.63	0.61	0.61	0.64	-0.56	-0.56	0.62	0.48	-0.39	-0.46		0.61	0.41	0.41	0.39	0.39	-0.62	0.62	-0.13	0.08
13	-0.97	0.97	1.00	1.00	0.96	-0.93	-0.93	0.96	0.82	-0.71	-0.74	0.61		0.73	0.77	0.80	0.73	-0.84	0.84	-0.26	0.11
14	-0.77	0.75	0.74	0.73	0.75	-0.89	-0.89	0.75	0.99	-0.98	-0.84	0.41	0.73		0.96	0.90	0.91	-0.37	0.37	-0.41	0.31
15	-0.81	0.79	0.78	0.77	0.75	-0.94	-0.94	0.76	0.95	-0.93	-0.79	0.41	0.77	0.96		0.94	0.89	-0.40	0.40	-0.36	0.26
16	-0.77	0.76	0.81	0.80	0.72	-0.89	-0.89	0.76	0.90	-0.88	-0.75	0.39	0.80	0.90	0.94		0.90	-0.39	0.39	-0.35	0.24
17	-0.73	0.72	0.74	0.73	0.71	-0.84	-0.84	0.81	0.90	-0.90	-0.79	0.39	0.73	0.91	0.89	0.90		-0.37	0.37	-0.41	0.28
18	0.83	-0.85	-0.83	-0.84	-0.87	0.66	0.66	-0.82	-0.50	0.35	0.51	-0.62	-0.84	-0.37	-0.40	-0.39	-0.37		-1.00	0.10	0.02
19	-0.83	0.85	0.83	0.84	0.87	-0.66	-0.66	0.82	0.50	-0.35	-0.51	0.62	0.84	0.37	0.40	0.39	0.37	-1.00		-0.10	-0.02
20	0.27	-0.26	-0.27	-0.26	-0.28	0.32	0.32	-0.31	-0.39	0.58	0.63	-0.13	-0.26	-0.41	-0.36	-0.35	-0.41	0.10	-0.10		-0.15
21	-0.13	0.12	0.11	0.11	0.13	-0.19	-0.19	0.14	0.28	-0.31	-0.53	0.08	0.11	0.31	0.26	0.24	0.28	0.02	-0.02	-0.15	

B	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1		-1.00	-0.92	-0.89	-0.94	0.93	0.93	-0.86	-0.81	0.60	0.39	-0.40	-0.89	-0.68	-0.78	-0.70	-0.60	0.57	-0.57	0.11	0.10
2	-1.00		0.92	0.90	0.94	-0.91	-0.91	0.86	0.79	-0.58	-0.38	0.40	0.90	0.65	0.75	0.67	0.58	-0.61	0.61	-0.10	-0.12
3	-0.92	0.92		0.93	0.86	-0.87	-0.87	0.86	0.74	-0.58	-0.37	0.36	0.93	0.63	0.73	0.81	0.63	-0.53	0.53	-0.14	-0.10
4	-0.89	0.90	0.93		0.86	-0.75	-0.75	0.81	0.65	-0.41	-0.29	0.38	1.00	0.47	0.56	0.62	0.46	-0.73	0.73	-0.04	-0.23
5	-0.94	0.94	0.86	0.86		-0.83	-0.83	0.88	0.83	-0.60	-0.40	0.42	0.86	0.68	0.65	0.58	0.57	-0.64	0.64	-0.10	-0.12
6	0.93	-0.91	-0.87	-0.75	-0.83		1.00	-0.81	-0.89	0.77	0.46	-0.34	-0.75	-0.84	-0.95	-0.85	-0.76	0.27	-0.27	0.19	-0.04
7	0.93	-0.91	-0.87	-0.75	-0.83	1.00		-0.81	-0.89	0.77	0.46	-0.34	-0.75	-0.84	-0.95	-0.85	-0.76	0.27	-0.27	0.19	-0.04
8	-0.86	0.86	0.86	0.81	0.88	-0.81	-0.81		0.79	-0.63	-0.41	0.38	0.81	0.69	0.69	0.68	0.80	-0.49	0.49	-0.16	-0.05
9	-0.81	0.79	0.74	0.65	0.83	-0.89	-0.89	0.79		-0.85	-0.53	0.31	0.65	0.95	0.87	0.77	0.81	-0.19	0.19	-0.19	0.09
10	0.60	-0.58	-0.58	-0.41	-0.60	0.77	0.77	-0.63	-0.85		0.69	-0.16	-0.41	-0.92	-0.84	-0.77	-0.81	-0.09	0.09	0.57	-0.28
11	0.39	-0.38	-0.37	-0.29	-0.40	0.46	0.46	-0.41	-0.53	0.69		-0.02	-0.29	-0.53	-0.47	-0.43	-0.47	0.00	0.00	0.64	-0.69
12	-0.40	0.40	0.36	0.38	0.42	-0.34	-0.34	0.38	0.31	-0.16	-0.02		0.38	0.24	0.25	0.20	0.21	-0.34	0.34	0.11	-0.16
13	-0.89	0.90	0.93	1.00	0.86	-0.75	-0.75	0.81	0.65	-0.41	-0.29	0.38		0.47	0.56	0.62	0.46	-0.73	0.73	-0.04	-0.23
14	-0.68	0.65	0.63	0.47	0.68	-0.84	-0.84	0.69	0.95	-0.92	-0.53	0.24	0.47		0.91	0.81	0.86	0.05	-0.05	-0.25	0.21
15	-0.78	0.75	0.73	0.56	0.65	-0.95	-0.95	0.69	0.87	-0.84	-0.47	0.25	0.56	0.91		0.90	0.82	0.00	0.00	-0.23	0.16
16	-0.70	0.67	0.81	0.62	0.58	-0.85	-0.85	0.68	0.77	-0.77	-0.43	0.20	0.62	0.81	0.90		0.81	0.00	0.00	-0.25	0.14
17	-0.60	0.58	0.63	0.46	0.57	-0.76	-0.76	0.80	0.81	-0.81	-0.47	0.21	0.46	0.86	0.82	0.81		0.06	-0.06	-0.27	0.19
18	0.57	-0.61	-0.53	-0.73	-0.64	0.27	0.27	-0.49	-0.19	-0.09	0.00	-0.34	-0.73	0.05	0.00	0.00	0.06		-1.00	-0.14	0.42
19	-0.57	0.61	0.53	0.73	0.64	-0.27	-0.27	0.49	0.19	0.09	0.00	0.34	0.73	-0.05	0.00	0.00	-0.06	-1.00		0.14	-0.42
20	0.11	-0.10	-0.14	-0.04	-0.10	0.19	0.19	-0.16	-0.19	0.57	0.64	0.11	-0.04	-0.25	-0.23	-0.25	-0.27	-0.14	0.14		-0.29
21	0.10	-0.12	-0.10	-0.23	-0.12	-0.04	-0.04	-0.05	0.09	-0.28	-0.69	-0.16	-0.23	0.21	0.16	0.14	0.19	0.42	-0.42	-0.29	

C	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1		-1.00	-1.00	-0.99	-0.99	0.97	0.97	-0.99	-0.88	0.77	0.56	-0.85	-0.99	-0.81	-0.86	-0.87	-0.84	0.92	-0.92	-0.74	-0.06
2	-1.00		1.00	0.99	0.99	-0.97	-0.97	0.99	0.88	-0.76	-0.55	0.84	0.99	0.80	0.85	0.86	0.82	-0.93	0.93	0.73	0.06
3	-1.00	1.00		0.99	0.99	-0.97	-0.97	0.99	0.88	-0.76	-0.54	0.84	0.99	0.81	0.86	0.88	0.84	-0.91	0.91	0.73	0.05
4	-0.99	0.99	0.99		0.97	-0.98	-0.98	0.98	0.91	-0.81	-0.56	0.86	1.00	0.85	0.89	0.91	0.87	-0.88	0.88	0.77	0.06
5	-0.99	0.99	0.99	0.97		-0.95	-0.95	0.99	0.87	-0.75	-0.56	0.84	0.97	0.79	0.82	0.82	0.81	-0.94	0.94	0.72	0.07
6	0.97	-0.97	-0.97	-0.98	-0.95		1.00	-0.98	-0.95	0.87	0.59	-0.89	-0.98	-0.91	-0.95	-0.95	-0.93	0.82	-0.82	-0.82	-0.09
7	0.97	-0.97	-0.97	-0.98	-0.95	1.00		-0.98	-0.95	0.87	0.59	-0.89	-0.98	-0.91	-0.95	-0.95	-0.93	0.82	-0.82	-0.82	-0.09
8	-0.99	0.99	0.99	0.98	0.99	-0.98	-0.98		0.91	-0.81	-0.58	0.88	0.98	0.85	0.87	0.88	0.88	-0.89	0.89	0.78	0.09
9	-0.88	0.88	0.88	0.91	0.87	-0.95	-0.95	0.91		-0.97	-0.67	0.91	0.91	0.98	0.97	0.95	0.97	-0.68	0.68	0.88	0.18
10	0.77	-0.76	-0.76	-0.81	-0.75	0.87	0.87	-0.81	-0.97		0.70	-0.88	-0.81	-0.99	-0.95	-0.92	-0.96	0.53	-0.53	-0.89	-0.24
11	0.56	-0.55	-0.54	-0.56	-0.56	0.59	0.59	-0.58	-0.67	0.70		-0.67	-0.56	-0.67	-0.62	-0.59	-0.65	0.45	-0.45	-0.69	-0.78
12	-0.85	0.84	0.84	0.86	0.84	-0.89	-0.89	0.88	0.91	-0.88	-0.67		0.86	0.90	0.88	0.87	0.90	-0.68	0.68	0.98	0.26
13	-0.99	0.99	0.99	1.00	0.97	-0.98	-0.98	0.98	0.91	-0.81	-0.56	0.86		0.85	0.89	0.91	0.87	-0.88	0.88	0.77	0.06
14	-0.81	0.80	0.81	0.85	0.79	-0.91	-0.91	0.85	0.98	-0.99	-0.67	0.90	0.85		0.97	0.95	0.98	-0.58	0.58	0.90	0.22
15	-0.86	0.85	0.86	0.89	0.82	-0.95	-0.95	0.87	0.97	-0.95	-0.62	0.88	0.89	0.97		0.99	0.98	-0.62	0.62	0.87	0.15
16	-0.87	0.86	0.88	0.91	0.82	-0.95	-0.95	0.88	0.95	-0.92	-0.59	0.87	0.91	0.95	0.99		0.96	-0.64	0.64	0.85	0.12
17	-0.84	0.82	0.84	0.87	0.81	-0.93	-0.93	0.88	0.97	-0.96	-0.65	0.90	0.87	0.98	0.98	0.96		-0.60	0.60	0.89	0.19
18	0.92	-0.93	-0.91	-0.88	-0.94	0.82	0.82	-0.89	-0.68	0.53	0.45	-0.68	-0.88	-0.58	-0.62	-0.64	-0.60		-1.00	-0.54	-0.02
19	-0.92	0.93	0.91	0.88	0.94	-0.82	-0.82	0.89	0.68	-0.53	-0.45	0.68	0.88	0.58	0.62	0.64	0.60	-1.00		0.54	0.02
20	-0.74	0.73	0.73	0.77	0.72	-0.82	-0.82	0.78	0.88	-0.89	-0.69	0.98	0.77	0.90	0.87	0.85	0.89	-0.54	0.54		0.32
21	-0.06	0.06	0.05	0.06	0.07	-0.09	-0.09	0.09	0.18	-0.24	-0.78	0.26	0.06	0.22	0.15	0.12	0.19	-0.02	0.02	0.32	

D	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1		-1.00	-0.92	-0.90	-0.93	0.91	0.91	-0.86	-0.66	0.47	0.35	-0.31	-0.90	-0.54	-0.68	-0.59	-0.52	0.71	-0.71	-0.18	-0.02
2	-1.00		0.92	0.89	0.93	-0.89	-0.89	0.86	0.64	-0.44	-0.34	0.30	0.89	0.51	0.65	0.56	0.49	-0.74	0.74	0.16	0.00
3	-0.92	0.92		0.99	0.82	-0.88	-0.88	0.82	0.64	-0.45	-0.38	0.33	0.99	0.54	0.70	0.76	0.55	-0.60	0.60	0.22	0.10
4	-0.90	0.89	0.99		0.79	-0.89	-0.89	0.80	0.66	-0.49	-0.38	0.33	1.00	0.57	0.74	0.81	0.58	-0.54	0.54	0.23	0.09
5	-0.93	0.93	0.82	0.79		-0.77	-0.76	0.87	0.66	-0.48	-0.36	0.32	0.79	0.53	0.50	0.41	0.47	-0.80	0.80	0.18	-0.01
6	0.91	-0.89	-0.88	-0.89	-0.77		1.00	-0.82	-0.82	0.68	0.46	-0.42	-0.89	-0.76	-0.92	-0.80	-0.72	0.39	-0.39	-0.34	-0.14
7	0.91	-0.89	-0.88	-0.89	-0.76	1.00		-0.82	-0.82	0.68	0.46	-0.42	-0.89	-0.76	-0.92	-0.80	-0.72	0.39	-0.39	-0.34	-0.14
8	-0.86	0.86	0.82	0.80	0.87	-0.82	-0.82		0.75	-0.58	-0.47	0.44	0.80	0.66	0.65	0.57	0.77	-0.57	0.57	0.33	0.13
9	-0.66	0.64	0.64	0.66	0.66	-0.82	-0.82	0.75		-0.93	-0.59	0.55	0.66	0.98	0.85	0.72	0.85	-0.14	0.14	0.51	0.22
10	0.47	-0.44	-0.45	-0.49	-0.48	0.68	0.68	-0.58	-0.93		0.61	-0.52	-0.49	-0.96	-0.78	-0.65	-0.79	-0.05	0.05	-0.53	-0.22
11	0.35	-0.34	-0.38	-0.38	-0.36	0.46	0.46	-0.47	-0.59	0.61		-0.55	-0.38	-0.60	-0.49	-0.46	-0.57	0.02	-0.02	-0.53	-0.81
12	-0.31	0.30	0.33	0.33	0.32	-0.42	-0.42	0.44	0.55	-0.52	-0.55		0.33	0.55	0.45	0.41	0.53	-0.02	0.02	0.98	0.39
13	-0.90	0.89	0.99	1.00	0.79	-0.89	-0.89	0.80	0.66	-0.49	-0.38	0.33		0.57	0.74	0.81	0.58	-0.54	0.54	0.23	0.09
14	-0.54	0.51	0.54	0.57	0.53	-0.76	-0.76	0.66	0.98	-0.96	-0.60	0.55	0.57		0.85	0.74	0.86	0.02	-0.02	0.55	0.27
15	-0.68	0.65	0.70	0.74	0.50	-0.92	-0.92	0.65	0.85	-0.78	-0.49	0.45	0.74	0.85		0.89	0.80	-0.04	0.04	0.44	0.22
16	-0.59	0.56	0.76	0.81	0.41	-0.80	-0.80	0.57	0.72	-0.65	-0.46	0.41	0.81	0.74	0.89		0.72	0.00	0.00	0.41	0.27
17	-0.52	0.49	0.55	0.58	0.47	-0.72	-0.72	0.77	0.85	-0.79	-0.57	0.53	0.58	0.86	0.80	0.72		0.01	-0.01	0.53	0.32
18	0.71	-0.74	-0.60	-0.54	-0.80	0.39	0.39	-0.57	-0.14	-0.05	0.02	-0.02	-0.54	0.02	-0.04	0.00	0.01		-1.00	0.16	0.22
19	-0.71	0.74	0.60	0.54	0.80	-0.39	-0.39	0.57	0.14	0.05	-0.02	0.02	0.54	-0.02	0.04	0.00	-0.01	-1.00		-0.16	-0.22
20	-0.18	0.16	0.22	0.23	0.18	-0.34	-0.34	0.33	0.51	-0.53	-0.53	0.98	0.23	0.55	0.44	0.41	0.53	0.16	-0.16		0.42
21	-0.02	0.00	0.10	0.09	-0.01	-0.14	-0.14	0.13	0.22	-0.22	-0.81	0.39	0.09	0.27	0.22	0.27	0.32	0.22	-0.22	0.42	

Table 4: Spearman correlation matrices for EHC (part A), SoNaR (part B), BNC (part C) and WSJ (part D). Correlations with an absolute value above 0.60 are displayed in bold.

measured in characters increases, so will the number of syllables per word. Furthermore, there is very little variance in the average number of characters per syllable and per text: we measured a standard deviation of 0.09 for EHC, 0.12 for SoNaR, 0.07 for BNC and 0.12 for WSJ.

Although word length and occurrence of words in a word frequency list are both assumed to indicate lexical complexity, we found no notable correlation between them for the Dutch corpora.

Sentence length The characteristics *avgpolysylsent* (13), *psw* (18) and *avgsentencelen* (19) are closely related to the number of words in each sentence. There is a perfect negative correlation between *psw* and *avgsentencelen*, because there exists a mathematical relationship between the two, defined by the equation $psw = 100/avgsentencelen$. In EHC (part A), SoNaR (part B) and BNC (part C), there is also a strong correlation of those variables with *avgpolysylsent*.

Type-token ratio The type-token ratio is a measure for lexical richness and does not seem related to any other variable. It is a variable in the *clib* formula. In order to construct that formula, a number of texts of almost equal length was used as a training corpus. Baayen (2008) mentions that type-token ratio strongly depends on text length. Unfortunately, the *clib* formula can therefore only be used for texts of the same length as in the training corpus.

3.3.2. Correlations between text characteristics and readability formulas

In general, for all corpora we observed moderate to strong correlations between readability scores (1 to 12) and text characteristics related to word length (13 to 17).

A variable related to word length in a readability formula clearly has a great influence on the readability scores calculated with that formula. Regardless of how word length is effectively measured, it is clear that a greater average word length is understood to indicate a more difficult text and a smaller average word length a more readable one. In that respect, an interesting phenomenon becomes apparent regarding the *cilt* and *clib* formulas. Those formulas should both yield higher scores for more difficult texts because their slope is positive (see table 2). However, for all corpora, even both EHC (part A) and SoNaR (part B), we see that the correspondence between *cilt* (10) and *clib* (11) scores on the one hand and word length (13 to 17) on the other is generally the opposite of what one would expect. Higher scores correspond with lower word length and vice versa. Supposing that word length is an indicator of text difficulty, simply the inverse of the *clib* and *cilt* formulas would be much better readability predictors than those formulas themselves.

3.3.3. Correlations between readability formulas

For Dutch, all examined readability formulas correlate strongly with superficial text characteristics that are closely related to word length.

The readability formulas are designed for different languages. It could therefore be expected that the formulas designed for other languages than Dutch are not applicable to a Dutch text corpus. However, strong correlations between readability formulas designed for English, Swedish and Dutch are presented in the correlation matrices in table 4. This observation can be explained by the fact that the formulas primarily make use of language-independent properties.

There are two language-dependent text characteristics which feature in readability formulas, viz. *freq77* (20 in EHC and SoNaR) in *clib* (11) and *cilt* (10) and *freq3000* (20 in BNC and WSJ) in *rgs* (12). Assuming that the other readability formulas (1 to 9) indicate the readability of a text, one would therefore expect higher correlations between them and *clib* and *cilt* for the Dutch corpora and between them and *rgs* for the English ones. However, this is not consistently borne out by the results.

3.3.4. Principal component analysis (PCA)

To further explain the results of the correlation matrices, we performed a principal component analysis to determine the number of *independent* text characteristics.

The results of the PCA are displayed in figure 1. For each corpus, two bar charts are shown. The charts on the left display the variances explained by the latent factors constituted of the readability formulas. The latent factors are vectors containing weights attached to each readability formula. For all corpora, the charts on the left can be interpreted in the same way. A single latent factor explains virtually all of the variance between the readability formulas. That means that if the readability scores are rescaled, they would be very similar, regardless of which formula is used. That confirms the results for the correlation matrices – if we disregard the slope of the readability formulas – where we observed strong and moderate correlations between all formulas.

The bar charts on the right are similar, but now the explained variances of the variables occurring in the readability formulas are displayed. Almost all of the variance is explained by at most five latent factors, which are now vectors consisting of weights for the variables. For the text characteristics, unlike for the readability formulas, it is meaningful to identify which ones contribute most to the most prominent latent factors. That identification is presented in table 5. As mentioned before, *freq77* was only taken into account for the Dutch and *freq3000* only for the English corpora.

3.3.5. Collinearity

In order to ascertain whether the methodology to construct the readability formulas is valid, we performed collinearity tests as proposed by Belsley et al. (1980). The result of such tests are condition numbers, which estimate the extent to which the variables occurring within the same formula are not independent. According to Baayen (2008), there is no appreciable collinearity if the condition number is between 0 and 6. Around 15, there is medium collinearity, and at or above 30, the collinearity is potentially harmful.

Table 6 lists the condition numbers acquired from the collinearity test. None of the condition numbers listed in table 6 are negligible, which means that for all formulas, multiple variables explain the same variance to a lesser or greater extent. Many of the condition numbers are greater

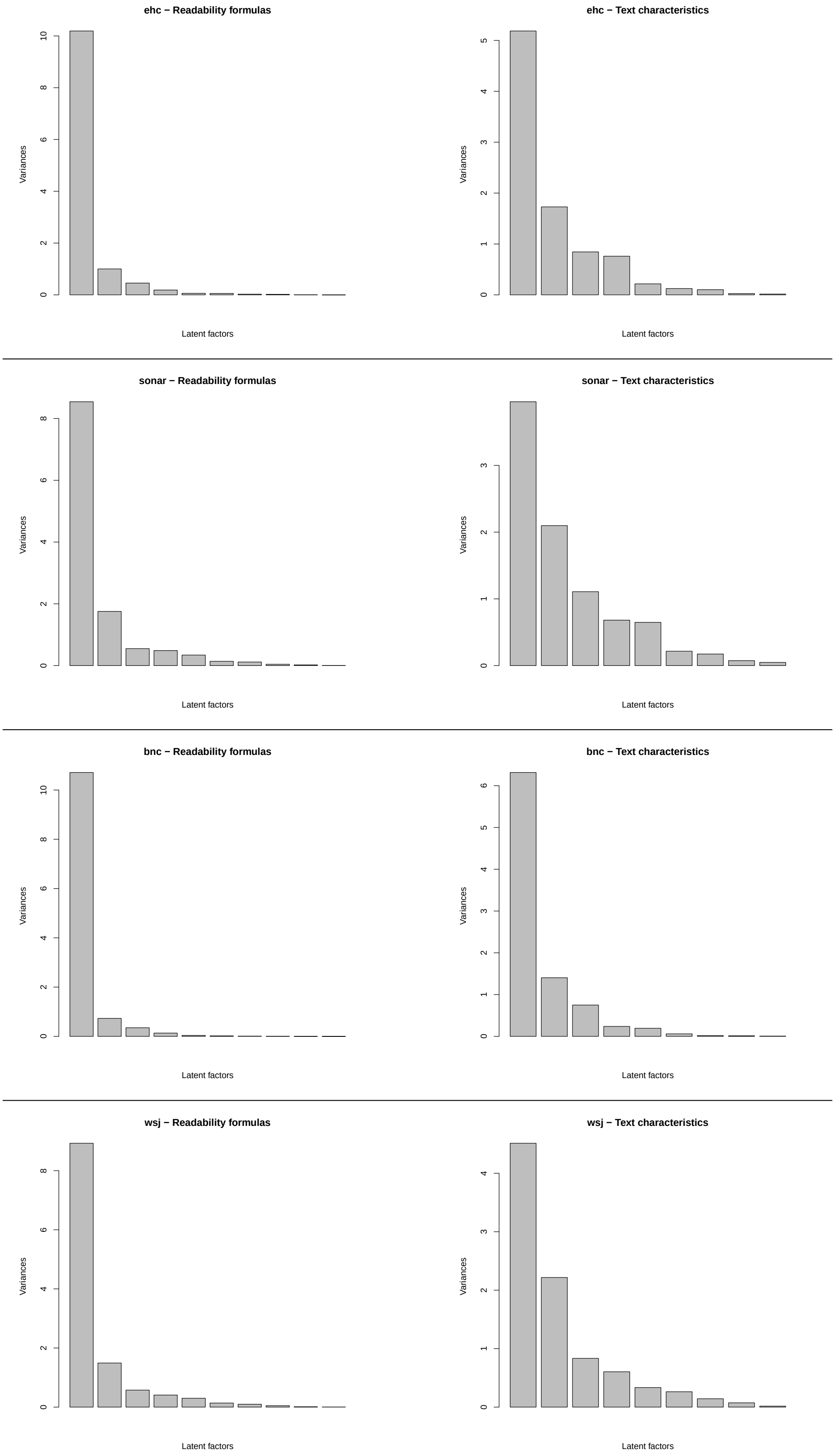

Figure 1: For each corpus, two bar charts showing the explained variances of principal components. For the charts on the left, the principal components are found by comparing the 12 readability formulas put forward in table 2. For those on the right, the variables used in the readability formulas were compared.

EHC	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**
avgpolysylsent	-0.38	-0.32	-0.02	0.01	-0.51
avgwordlen	-0.40	0.22	-0.04	0.15	0.09
avgnumsyl	-0.40	0.18	-0.06	0.22	0.04
ppolysylword	-0.40	0.16	-0.07	0.23	-0.07
ratiolongword	-0.39	0.20	-0.05	0.11	0.14
psw	0.28	0.48	-0.17	0.15	-0.75
avgsentencelen	-0.26	-0.57	0.08	-0.19	-0.33
ttr	-0.13	0.32	0.87	-0.30	-0.11
freq77	0.19	-0.26	0.42	0.83	0.00

SoNaR	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**
avgpolysylsent	0.14	-0.61	0.20	-0.19	0.07
avgwordlen	0.46	-0.00	-0.14	0.03	-0.11
avgnumsyl	0.46	-0.04	-0.21	0.03	-0.14
ppolysylword	0.45	-0.04	-0.20	0.02	-0.04
ratiolongword	0.45	0.03	-0.12	-0.08	0.08
psw	0.20	0.36	0.26	-0.44	0.71
avgsentencelen	0.01	-0.64	0.23	-0.19	0.11
ttr	0.15	0.25	0.61	-0.37	-0.62
freq77	-0.23	-0.02	-0.57	-0.76	-0.17

BNC	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**
avgpolysylsent	-0.37	0.20	0.03	-0.38	0.20
avgwordlen	-0.37	-0.18	-0.15	0.12	-0.02
avgnumsyl	-0.38	-0.10	-0.17	0.15	0.23
ppolysylword	-0.38	-0.07	-0.18	0.09	0.39
ratiolongword	-0.38	-0.16	-0.15	0.07	0.04
psw	0.24	-0.51	-0.41	-0.68	0.12
avgsentencelen	-0.29	0.42	0.34	-0.56	-0.04
ttr	-0.06	-0.61	0.76	-0.00	0.17
freq3000	-0.34	-0.24	-0.05	-0.08	-0.83

WSJ	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**
avgpolysylsent	0.37	-0.32	0.02	-0.33	0.34
avgwordlen	0.42	0.09	-0.15	0.20	-0.35
avgnumsyl	0.43	0.00	-0.26	-0.06	0.03
ppolysylword	0.42	0.02	-0.20	-0.26	0.35
ratiolongword	0.42	0.08	-0.07	0.14	-0.50
psw	-0.07	0.58	-0.21	-0.27	0.29
avgsentencelen	0.07	-0.60	0.31	-0.09	0.00
ttr	0.17	0.35	0.70	-0.49	-0.26
freq3000	0.28	0.22	0.46	0.65	0.47

Table 5: Identification of the first 5 principal components in the analysis of the variables used in the readability formulas, for each text corpus.

Formula	**Text characteristics**	**Condition number**			
		EHC	SoNaR	BNC	WSJ
flesch *kincaid* *douma* *brouwer*	*avgsentencelen* *avgnumsyl*	21.7	24.4	33.0	33.8
clib	*freq77* *avgwordlen* *ttr* *psw*	60.1	44.5	–	–
cilt	*freq77* *avgwordlen*	44.0	34.5	–	–
lix	*avgsentencelen* *ratiolongword*	9.7	10.6	11.5	15.4
rgs	*freq3000* *avgsentencelen*	–	–	12.4	17.8
cli *ari*	*avgwordlen* *avgsentencelen*	23.0	25.7	38.0	41.8
fog	*avgsentencelen* *ppolysylword*	7.8	8.9	9.9	12.6

Table 6: Collinearity of the variables that occur together in at least one readability formula. The condition numbers are calculated as proposed by Belsley (1980). Formulas containing language-specific text characteristics are only tested for the language the formula is constructed for.

than 30, which indicates that the readability formula in question can probably not be used for texts in the respective corpus. The condition numbers for *cilt* and *clib* are very high, indicating considerable collinearity between the variables that constitute those formulas. Together with the observation that the correlations for *cilt* and *clib* are the opposite of what was expected (see paragraph 3.3.2.), the probability that *cilt* and *clib* are usable seems very low.

4. Discussion

The number of different features used in the readability formulas is limited, especially so because the results of principal component analyses show that some of the features used to predict readability strongly overlap. Also, there is a strong correspondence between the results of the readability formulas. Particularly remarkable are the strong correspondences between the readability formulas intended for different languages, as we established in two different ways. However, it does not seem reasonable to assume that a formula consisting of strictly language-independent text characteristics can accurately predict the readability of a text in a natural language. For instance, in Dutch, as opposed to English, compounds are generally written as one word. It would seem then, that the feature of word length is a less valid indicator for word difficulty in Dutch than in English. Therefore, it could be assumed that word length for decompounded text would give a more accurate indication of word difficulty for Dutch. This effect may also explain the weaker correlation between occurrence in a word frequency list and word length for Dutch.

Apart from language-specific features, thanks to the advances in Natural Language Processing, we now have an enormous range of linguistic features at our disposal; thanks to the advances in Machine Learning, we now have the ability of processing vast amounts of information; and thanks to the advances in language psychology, we are aware of a number of deeper-lying text characteristics that have an influence on readability. For instance, Graesser et al. (2004) highlight the effects of text cohesion on comprehension. We hypothesize that a better readability prediction can be achieved by means of a greater range of features.

The methodology that was used to come to the readability formulas does not seem valid in most cases, which was borne out by a validity test of the preconditions on which the methodology to derive readability formulas is founded against large corpora. For some formulas, the invalidity is even obvious, given the collinearity of the text characteristics used in the formulas. For *cilt* and *clib*, we can even show that their results are incorrect. In order to show that those formulas can still be used for their intended purpose, namely assessing readability within the specific domain of children's literature, those preconditions would have to be checked against a *large* corpus of texts in that same domain. It is very important that the construction of state-of-the-art readability prediction systems will be based on a clean methodology, of which the preconditions must be validated against large corpora whenever possible. Research that aims to construct a new readability prediction system should therefore be embedded in corpus research.

Another condition for a successful methodology is that each text in the corpus that is used to compile a readability prediction system must be evaluated by means of at least two different kinds of readability assessment. It is clear that that is necessary to disengage readability of a text from its assessment, because otherwise the system will only be trained to predict the assessment results instead of the readability of the text itself. A gold standard corpus must be composed in order to develop a new corpus-based readability prediction system.

In order to compile such a gold standard, we have created two web applications designed to collect readability assessments for Dutch texts: one that is intended exclusively for Dutch language experts[3], and one that is open to the general public[4]. In the former, the experts are asked to rank a number of texts according to their readability, while users of the latter are presented with two texts and asked to select the most difficult one. In this way, we aspire to employ the opinions of actual readers to score and compare readability. The idea behind the design of these web applications is that the readability of a text can be conceptualized as the extent to which the text is perceived to be readable by the community of language users. The evaluation of the data gathered with the applications will result in a new gold standard, which will in turn be used to train new readability prediction systems.

5. References

R. H. Baayen, R. Piepenbrock, and L. Gulikers. 1995. The CELEX Lexical Database (CD-ROM). Linguistic Data Consortium.

R. H. Baayen. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R.* Cambridge University Press.

David A. Belsley, Edwin Kuh, and Roy E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, August.

Carl-Hugo Björnsson. 1968. *Läsbarhet*. Almqvist and Wiksell, Stockholm.

British National Corpus Consortium. 2000. British national corpus world edition. CD-ROM.

R. H. M. Brouwer. 1963. Onderzoek naar de leesmoeilijkheden van Nederlands proza. *Pedagogische Studiën*, 40:454–464.

Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284.

Walter Daelemans and Antal van den Bosch. 1992. Generalization performance of backpropagation learning on a syllabification task. In M.F.J. Drossaers and A. Nijholt, editors, *Connectionism and Natural Language Processing. Proceedings Third Twente Workshop on Language Technology*, pages 27–38.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational research bulletin*, 27:11–20.

W.H. Douma. 1960. De leesbaarheid van landbouwbladen: een onderzoek naar en een toepassing van leesbaarheidsformules. *Bulletin*, 17.

William H. DuBay. 2004. *The Principles of Readability*. Impact Information.

1989. Eindhoven corpus. txt. VU-versie.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, June.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments and Computers*, 36:193–202.

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

J. Peter Kincaid, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Research branch report RBR-8-75, Naval Technical Training Command Millington Tenn Research Branch, Springfield, Virginia.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19:313–330.

G. Harry McLaughlin. 1969. SMOG grading – a new readability formula. *Journal of Reading*, pages 639–646.

G. Harry McLaughlin. 1974. Temptations of the flesch. *Instructional Science*, 2(4):367–383, January.

J.P. Royston. 1982. An Extension of Shapiro and Wilk's *W* Test for Normality to Large Samples. *Applied Statistics*, 31(2):115–124.

Ineke Schuurman, Veronique Hoste, and Paola Monachesi. 2009. Cultivating Trees: Adding Several Semantic Layers to the Lassy Treebank in SoNaR. In *Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories*, Groningen, The Netherlands.

R. J. Senter and E. A. Smith. 1967. Automated readability index. Technical Report AMRLTR-66-220, University of Cincinnati, Cincinnati, Ohio.

Gerrit Staphorsius. 1994. *Leesbaarheid en leesvaardigheid, De ontwikkeling van een domeingericht meetinstrument.* Cito, Arnhem.

StatSoft Inc. 2007. Electronic statistics textbook. WEB. http://www.statsoft.com/textbook/stathome.html.

[3] The application itself, which is password protected, is located at `http://lt3.hogent.be/hendi/expert-readers`. An animated demo is available from `http://www.youtube.com/watch?v=iSXBQIlwslo`

[4] `http://lt3.hogent.be/hendi/sort`