Computational Linguistics in the Netherlands Journal 9 (2019) 81–100 Submitted 09/2019; Published 12/2019

Linguistic Proxies of Readability: Comparing Easy-to-Read and regular newspaper Dutch

Vincent Vandeghinste* VINCENT.VANDEGHINSTE@IVDNT.ORG
Bram Bulté** BRAM.BULTE@CCL.KULEUVEN.BE

* *Instituut voor de Nederlandse Taal, Leiden, Netherlands*

** *Centre for Computational Linguistics, KU Leuven, Belgium*

Abstract

The aim of this study is to identify linguistic proxies of readability in Dutch, i.e. those linguistic features that define text as being easy-to-read. To this end, we compare the Wablieft corpus (Vandeghinste et al. 2019) (Flemish easy-to-read newspaper archives) to articles that appeared in the regular Flemish newspaper *De Standaard*, using a wide range of lexical, syntactic and readability metrics. We test which of these metrics has the highest effect size and which combinations of metrics work best in a classification task predicting whether articles belong to *Wablieft* or *De Standaard*.

The results indicate that the best linguistic proxy for readability is (not surprisingly) the average number of words per sentence. Traditional reading metrics score well, although the combination of the parameters constituting these metrics score better in logistic regression than the original metrics.

1. Introduction

Many texts are too difficult to read for many people. One out of ten people in Belgium are considered illiterate,[1] meaning they have trouble reading and writing. Even more people are low literate or nearly functionally illiterate. They have trouble reading text on paper and on web sites. According to the definition of the United Nations, *"a person is functionally illiterate who cannot engage in all those activities in which literacy is required for effective functioning of his group and community and also for enabling him to continue to use reading, writing and calculation for his own and the community's development"* (United Nations 1984).

One can try to alleviate this problem by means of adult education aimed at promoting literacy, but such an approach will always miss out part of the targeted population. Another approach is to address text difficulty. Certain texts are difficult to read, for example, because they contain too many difficult words and/or complicated sentences. For this reason, many governments and institutions, including the Flemish government, have adopted policies and programs to advocate clear and simple language use.[2]

The Wablieft organisation[3] addresses the issue of text difficulty and readability on two sides.

1. On the reading side, people who want to read easy texts are offered the Wablieft newspaper. This weekly newspaper, established in 1989, is written in easy-to-read language, and currently has more than 45,000 readers. An online archive of articles published since 2009 is available on the organisation's website, and has been made available for research as a corpus, as described in Vandeghinste et al. (2019). This corpus can be downloaded from the *Taalmaterialen* section of the website of the Dutch Language Institute,[4] or queried with the GrETEL and PaQu syntactic

1. https://www.belgium.be/nl/Leren/permanente_vorming/alfabetisering
2. https://overheid.vlaanderen.be/heerlijk-helder
3. http://www.wablieft.be/
4. https://ivdnt.org/downloads/taalmaterialen/tstc-wablieft-corpus-1-1

search engines (Augustinus et al. 2017, Odijk et al. 2017).[5] It will soon be made searchable as part of the *Corpus Hedendaags Nederlands*.[6] Recently, a second, even easier-to-read newspaper was started, called Wablieft Start, but this was not included in this study.

2. On the writing side, the Wablieft organisation provides training sessions for people and organisations that want to publish easily readable texts. They also offer, against payment, to rewrite texts in clear language.

The Wablieft corpus provides an opportunity to study actual easy-to-read texts. Since these texts are written by professional, specialised authors, with the explicit purpose of being easy-to-read, the corpus can be considered a *gold standard* in this respect. In the present paper we investigate linguistic properties of texts in the Wablieft corpus by comparing them with regular newspaper articles taken from the Flemish newspaper *De Standaard*,[7] and test whether using these properties we can build a classifier that is able to discriminate between *regular* and *easy-to-read* articles. Such an analysis can be informative for research on automated text simplification, in the context of which the assessment of text readability or difficulty is often included as a first step (Bulté et al. 2018). Additionally, automated readability scoring has potential applications in the field of second language learning (e.g. for the purpose of selecting texts suited for learners at different proficiency levels) as well as for persons with low literacy.

For the purpose of this article, we define readability as the relative ease or difficulty with which a text is or can be read. We only focus on textual properties here, thus disregarding other aspects such as fonts and layout. Even though readability is a (potentially individually, temporally and situationally variable) subjective concept, having to do with cognitive processing load, our operationalisation is based on a more stable, objective interpretation of the term (i.e. some texts are, generally speaking, easier to read than others). Our aim, then, is to find objective, formal, structural or other features of texts that are related to the subjective notion of readability or, to put it differently, to find out which (linguistic) features differ between *easy-to-read* and *regular* texts.

In section 2 we present related work. Section 3 describes our methodology, while section 4 describes the results. Section 5 concludes and describes plans for future work.

2. Related work

Plenty of guidelines for writing easy-to-read text are available, such as those written by Inclusion-Europe (2009) or the guidelines on the Wablieft website. We present some in more detail in section 2.1. Section 2.2 presents available easy-to-read corpora, and in section 2.3 related work concerning readability prediction and classification is discussed.

2.1 Guidelines for writing easy-to-read text

There are a number of approaches towards writing text while addressing the issue of limited literacy. Some texts are written with the general aim of being *easy-to-read*, others have children or second language learners as their target group, and yet other texts aim at people with cognitive disabilities, so there is a large spectrum of possible target users.

The guidelines on the Wablieft website can be summarised as: "*address your readers, use short sentences (±15 words), and use everyday language*".

More extensive standards for writing easy-to-read text are described and illustrated in Inclusion-Europe (2009), which contains instructions on how to produce easy-to-read documents. The language-related instructions target different aspects of writing at the lexical, syntactic and discourse level.

5. `http://gretel.ccl.kuleuven.be/` and `https://paqu.let.rug.nl:8068/` respectively.
6. `http://corpushedendaagsnederlands.inl.nl/`
7. The articles from *De Standaard* cannot be made publicly available due to publication permissions, but they form part of the *Corpus Hedendaags Nederlands*

At the lexical level, writers are instructed to use words that people know and that are easy to understand, or to explain difficult words. Lexical variety is discouraged, as writers should be consistent with the words that they use and avoid introducing new words to refer to already introduced concepts. Also metaphors, loan words, abbreviations, long numbers and percentages are to be avoided. With regard to syntax, short, active sentences are preferred, in which the reader should be addressed directly. Sentences should also be affirmative. Finally, at the level of discourse, writers are instructed to structure the information in a simple way and to organise the text by topic. Repeating information is not discouraged.

2.2 Easy-to-read corpora

Probably the most well-known initiative with respect to easy-to-read text is Wikipedia Simple English,[8] which is written in so-called *simple English*. Amongst other things, its authors are instructed to use only the 1000 most frequent words of English. Other Wiki-initiatives are aiming at kids, such as Wikikids[9] for Dutch-speaking children and Vikidia[10] for kids speaking French, Italian, Spanish, English, Basque, Catalan, German, Russian, Greek and Sicilian.

In an academic context, a number of available easy-to-read corpora have been described and used, such as the Swedish LäSBarT corpus (Mühlenbock 2009), a corpus for Brazilian Portuguese (Aluísio et al. 2008), the French CLEAR medical corpus (Grabar and Cardon 2018), and a very small (227 sentences) corpus for Basque (Gonzalez-Dios et al. 2018). There exist some monolingual comparable corpora, in which *regular* text is aligned with its *easy-to-read* variant. Alignment can be at the text level, the paragraph level or the sentence level. A list of English comparable corpora with an easy-to-read side can be found in Yaneva (2015), and also for French (Cardon and Grabar 2018) and Brazilian Portuguese (de Medeiros Caseli et al. 2009) there have been efforts to create such a comparable corpus.

Very recently, Naderi et al. (2019) presented a dataset containing 1000 German sentences taken from 23 Wikipedia articles to be used for developing text-complexity predictor models and automatic text simplification. The dataset includes subjective evaluations of different text-complexity aspects provided by German learners. In addition, it contains 250 sentences that were manually simplified by native speakers and subjective assessment of these simplified sentences by target users.

Apart from the Wablieft corpus (Vandeghinste et al. 2019), we are not aware of any such efforts for Dutch, although the Dutch data in the CHILDES project might be worth mentioning (MacWhinney 2000), as well as the JASMIN speech corpus, consisting of recordings of Dutch speech by young people, non-native speakers, and elderly people (Cucchiarini et al. 2008). These two projects recorded supposedly easy *active* speech, whereas the Wablieft corpus contains texts focusing on the *passive* language knowledge of the target users.

Also worth mentioning is the corpus of 105 Dutch texts with readability judgements collected by De Clercq et al. (2013). Even though this is not an easy-to-read corpus, the readability assessments provided by experts and the crowd (De Clercq et al. 2014) allow for a categorisation of the texts according to their perceived readability. The lexical resource NT2Lex described in Tack et al. (2018) is a CEFR-graded[11] lexicon based on a corpus of texts targeted at Dutch language learners, which consists of CEFR-graded subcorpora. Unfortunately this corpus is not available.

2.3 Automated assessment of readability

Automated readability assessment has a long tradition dating back to the readability formulas developed in the early 20th century (Flesch 1948). These formulas were intended to provide an

8. `https://simple.wikipedia.org/wiki/Main_Page`
9. `http://www.wikikids.nl`
10. `http://www.vikidia.org`
11. Common European Framework of Reference for Languages

objective, quantitative evaluation of how easy or difficult to read texts are. Readability formulas have also been developed for analysing Dutch texts specifically (Brouwer 1963, Staphorsius 1994).

Since the turn of the century, readability prediction has seen a move away from these simple formulas towards data-driven machine learning approaches, that involve increasingly more and more complex features. Classifiers have been built using a wide range of approaches, such as naive Bayes (Collins-Thompson and Callan 2004), logistic regression (François 2009) and support vector machines (Schwarm and Ostendorf 2005, Larsson 2006). The features that are included in these classifiers range from surface-level word, sentence and text metrics (such as average length in terms of number of characters) to more complex features based on syntactic parse trees, language models and morphological features (Hancke et al. 2012, Dell'Orletta et al. 2011).

T-scan, a tool developed to automatically assess various textual features, includes some measures related to Dutch text difficulty and readability (Pander Maat et al. 2014). De Clercq and Hoste (2016) tested a more advanced machine learning approach to Dutch readability prediction using support vector machines, with a wide range of linguistic and textual features as input. The corpus they used to train their classifier was compiled by means of pairwise comparisons of 105 texts in terms of their readability.

3. Method

Section 3.1 describes the data sets we used. Section 3.2 describes the linguistic features or metrics under investigation. Section 3.3 explains the statistics used in the comparison of the two corpora, and section 3.4 describes the classifiers we trained.

3.1 Data

In order to find out which linguistic proxies reflect in what respect easy-to-read text differs from *regular* text, we compare the Wablieft data with the De Standaard 2010 data.

We use the articles in the Wablieft corpus (Vandeghinste et al. 2019) as our *easy-to-read* corpus, and as a control corpus, we use all the articles published in *De Standaard* year 2010. These data are part of the Corpus Hedendaags Nederlands (CHN), which we can unfortunately not make available for download due to copyright reasons, but which is available as a CLARIN infrastructure for online querying through `http://corpushedendaagsnederlands.inl.nl/`. Table 1 presents some statistics about the two corpora.

	Articles	**Sentences**	**Tokens**
Before Preprocessing			
Wablieft	12,683	256,729	2,074,491
De Standaard	34,520	905,875	14,745,751
After Preprocessing			
Wablieft	12,665	256,226	2,072,945
De Standaard	31,140	869,260	11,932,643

Table 1: Size of the data sets. The numbers presented under **Before Preprocessing** are the sizes of the raw data sets. Those in the **After Preprocessing** section are the sizes of the data sets after cleaning and filtering.

Both corpora are preprocessed by the Alpino parser (van Noord 2006). As a tokeniser we use the built-in tokeniser from Alpino.

Sentence detection for the Wablieft corpus is done semi-automatically. For *De Standaard* it is done fully automatically, based on punctuation. It is clear that the De Standaard data contain numerous sentence detection mistakes, especially in those cases where parts which should have been

detected as separate sentences, do not end in a full stop. This happens mainly in headings. We hope to compensate for this by the sheer number of sentences and articles in one year of daily newspaper processing, and by filtering out the worst cases.[12] We apply a number of filters to automatically clean the two corpora. These filters operate both at the article and the sentence level. At the article level, cleaning consists of excluding all articles with less than five sentences (on the assumption that these articles contain too little information), as well as articles with on average over 100 words per finite verb. This second filter is imposed in order to leave out 'non-language' articles such as stock market results, TV programme listings or sports results, since these can be considered to contain insufficient linguistic data. We believe that by putting the threshold at 100 words per finite verb, there is virtually no risk of also excluding 'real' newspaper articles.

At the level of individual sentences, cleaning consists of deleting those sentences with less than four characters, because these usually constitute the initials of the journalist in *De Standaard* data. Sentences that contain more than 20% words containing a digit, and which are longer than 200 characters, are skipped. So are sentences with a ratio of proper names (according to the parser) of more than 40%, no verbs and more than 100 characters long. Manual spot checks revealed that these filters only filter out non-linguistic data, as well as some cases where sentence boundary identification has failed. No 'regular' sentences were filtered out. The exact statistics of the corpora, after filtering, are presented in Table 1.

3.2 Linguistic features and readability metrics

We calculate a broad selection of metrics, ranging from lexical measures to more complex (morpho-) syntactic metrics and traditional readability formulas, to compare our two corpora. These features are also used in the classifiers (see section 3.4). Some of these measures use purely formal text criteria and rely on the identification of basic linguistic units (e.g. characters, words or sentences), whereas others require deeper syntactic parsing and/or part-of-speech labelling. Certain measures also use (text-)external information sources such as frequency lists or the results of psycholinguistic studies to assess word difficulty.

The lexical metrics are *words per text (w/txt)*,[13] *characters per word (c/w)*, *syllables per word (syl/w)*, *long word ratio (lwr)*, *type token ration (ttr)*, *Guiraud index (gi)*, *lexical density (ld)*, *frequent word ratio (fwr)*, *age of acquisition (aoa)* and *concreteness (conc)*. Words are operationalised as all tokens which are not tagged as punctuation. To determine the number of syllables per word, we use the rule-based syllable counter described in Vandeghinste and Pan (2004). The long word ratio is calculated as the amount of words that contain more than three syllables divided by the total amount of words. The type token ratio is the number of word types divided by the number of tokens in the text. The Guiraud index of lexical diversity (Guiraud 1954), which is a lexical type token ratio corrected for text length, is calculated as

$$gi = \frac{wordtypes}{\sqrt{tokens}} \tag{1}$$

Lexical density refers to the proportion of content words in a text, and is operationalised by dividing the number of content words (i.e. all nouns, adjective, adverbs and verbs, apart from the verbs which lemmatise to *hebben, zijn, worden* and *zullen*) by the total number of words. The frequent word ratio is calculated as the ratio of words that appear in the list of 77% most frequent words in the frequency list of the SONAR corpus (Oostdijk et al. 2013). This metric is also used in the CLIB and CILT formulas (Equations 4 and 5). Age of acquisition and concreteness are two

12. In future work, we could exclude headings from the analysis, as this information is included in the TEI file formats we received. For this paper, however, we realised this too late to redo the analyses.
13. One can dispute whether this is really a *lexical* metric, but it surely does not belong in the other categories. We added this metric because it provides a good descriptive statistic, as there are considerable differences in text length.

psycholinguistic measures based on data provided by Brysbaert et al. (2014). Aoa refers to the average age at which children learn a word, which can be seen as an indication of word difficulty. Concreteness evaluates the degree to which a concept denoted by a word refers to a perceptible entity, as perceived by native speakers. For both measures, we calculate average values for all content words in a text. Words that do not occur in the list with norms for 30,000 Dutch words, are given the maximum aoa as it occurs in the data. Words that do not occur in the list with norms for concreteness are not used in the calculation of average conc.

The morpho-syntactic metrics target the occurrence and relative proportion of specific word categories (De Clercq and Hoste 2016, Feng et al. 2010). We calculate the frequency of occurrence of the main part-of-speech labels (verbs (V), adjectives (ADJ), nouns (N), numerals (NUM), adverbs (ADV), punctuation signs (PUN), special tokens (SPEC), prepositions (PRP), pronouns (PRN), articles (ART), conjunctions (CONJ)[14]) and the ratio of each of these parts of speech, i.e. the relative frequency of the part-of-speech over all the tokens in the text, abbreviated as the tag abbreviation + r (e.g. Vr). We also include the number of finite verbs (finV).

The syntactic metrics consist of *words per sentence (w/sen)*, *number of clauses (cl)*, *clause length (cllen)*, *number of subclauses (subc)*, *subclauses per clause (subcl/cl)*, *subclause length (sublen)*, *words per finite verb (w/finV)*, *noun phrase length (NPlen)*, *dependents per head (d/h)* and *tree depth (depth)*. These measures either target the length or number of different syntactic units or the degree of embedding. For clause length, we count the following Alpino categories as clauses: `smain, sv1, cp, svan, rel, whrel, whq, whsub`. The length of a clause is taken as the subtraction of the value of the `end` feature of the clause with the value of the `begin` feature. For subclauses per clause we divide the number of subordinate clauses by the total number of clauses. As subordinate clauses, we use clauses which were given the following category labels by Alpino: `cp, svan, rel, whrel, whq, whsub`. For subclause length, we use only the subordinate clauses and use a similar calculation as for clause length. For words per finite verb, we divide the text length by the number of finite verbs. For dependents per head, we count, for each head in the Alpino trees, the number of siblings. For tree depth, we count the maximum depth of each syntactic tree.

The readability metrics are *Flesch, Flesch Douma, CLIB, CILT* and *Leesindex A*. Flesch is one of the oldest readability formulas, developed in 1948 (Flesch 1948). Its formula is presented in Equation 2, where I is the resulting Index value. The higher I, the easier the text.

$$I = 206.835 - (1.015 \times w/sen) - (84.6 \times syl/w) \tag{2}$$

Flesch Douma is the Dutch equivalent for Flesch Reading Ease (Douma 1960). The formula is

$$I = 206.83 - (0.93 \times syl/w) - (77 \times w/sen) \tag{3}$$

CLIB, CILT and Leesindex A were developed for assessing the readability of texts in a primary school context. CLIB is the Cito readability index for basic education (Leesbaarheidsindex voor het Basisonderwijs), developed for Dutch (Staphorsius 1994). The formula is presented in Equation 4. A CLIB score lower than 74 indicates that a text is suitable for primary education.

$$CLIB = 46 - (6.603 \times c/w) + (0.474 \times fwr) - (36.5 \times ttr) + (1.425 \times w/sen) \tag{4}$$

CILT is the Cito readability index for technical reading (Cito Leesindex Technisch Lezen) (Staphorsius 1994). It is calculated as follows:

$$CILT = 114.49 + 0.28 \times fwr - 12.33 \times c/w \tag{5}$$

14. The part-of-speech labels are provided by the parser, and are formed according to the CGN-tagset (Van Eynde 2004).

This index is used for determining AVI levels, i.e. the levels of reading proficiency required to read a certain text, as used in primary education. The Leesindex A (Brouwer 1963) is calculated according to formula 6. *Regular* text should score between 50 and 75, and a higher score indicates an easier text. This is the index that was used for AVI calculations before 1994 and the development of CLIB.

$$A = 195 - 2 \times w/sen - 66.7 \times syl/w \tag{6}$$

3.3 Comparative analyses

We report descriptive statistics (mean, median, standard deviation) per linguistic feature and metric to describe the two corpora, and compare them using independent samples t-tests with Cohen's d as a measure of effect size. We rely on the value of d to determine which metrics are best able to discriminate between both corpora or, put differently, to ascertain which linguistic features and metrics characterise easy-to-read text, when compared to regular text.[15]

We also present charts showing the frequency distributions of the scores for each measure, calculated once for each corpus (see Figure 1 and appendix A). To this end, we divided the scores in different bins and counted how many articles occur in each bin. This presents us with a visual resumé showing how large the overlap is between the different distributions.

3.4 Automated text classification

We build a series of binary classifiers with the aim of predicting whether a text is easy-to-read (i.e. belongs to the Wablieft corpus), or not (i.e. belongs to De Standaard). For this purpose, we take a random selection of 12,000 articles from each of the corpora, ensuring a balanced distribution. The data are then randomly divided into a training (90%) and test set (10%).

We test two types of classifiers. The first type are **logistic regression binary classifiers**, implemented using the `sklearn Logistic Regression` Python3 library. We start an A* search (Dechter and Pearl 1985) for the best feature combination by training the model on each feature separately. We then sort the models in descending order based on their accuracy, and expand the best scoring hypothesis by adding a second feature. These results are added to the stack of non-expanded feature sets, and the best scoring one is expanded again. This algorithm was run until no further improvements were found. We also applied A* with a correction on number of features, to explore other parts of the search space, left unattended by the regular A*. In addition, we tested all possible permutations up to five features. These alternate searches occasionally led to improved results on best scoring feature sets for a given number of features. Results are presented in section 4.

The second type are **multilayer perceptron binary classifiers**, implemented using the `keras` Python3 library with up to four hidden layers of 2 to 100 dimensions each, 30% dropout, L2 regularisation, and a rectified linear activation function, where the input layer consists of one input node per feature. For initial testing we tried different batch sizes (10 up to 12000) and different numbers of epochs (10 and 100). Eventually, we decided to run an A* search, similar to the one for logistic regression. Apart from the features, we also included different batch sizes (10 and 30) as well as different numbers of epochs (10 and 30) in the A* search space in order to explore whether these would yield better results. Because these neural models work with random initialisations, different training sessions on the same data can lead to different results, for example due to local minima encountered by the gradient descent algorithm. Therefore, we trained each network 30 times and used the average prediction accuracy of these 30 runs.

15. We do not use these parametric tests here to make claims about the statistical significance of observed differences (considering the size of the data set we use, too high statistical power would render such an analysis meaningless), but rather as an indication of how discriminating individual features are (i.e. we are only interested in an estimate of the size of the effect).

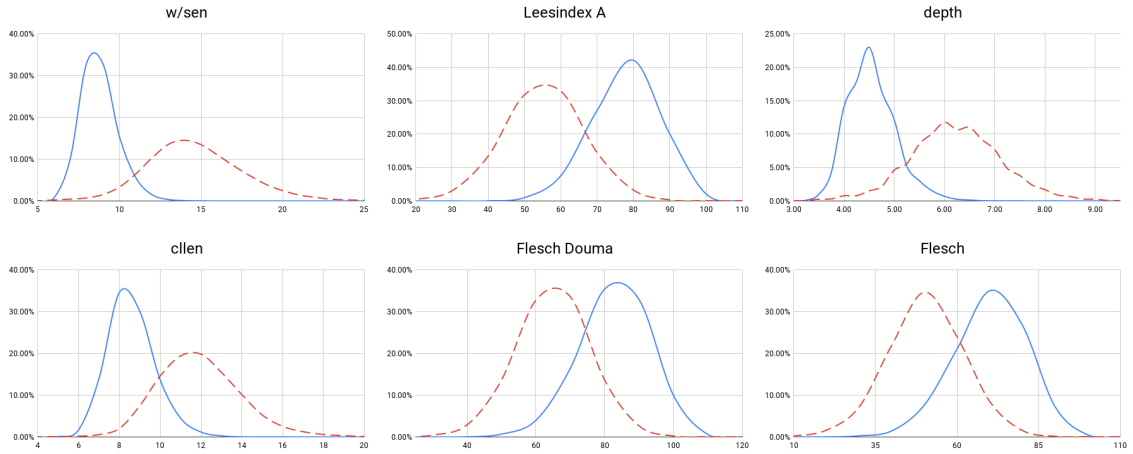

Figure 1: Distributions of the six best-scoring features. Continuous blue lines represent the Wablieft distribution, dotted red lines the De Standaard distribution.

4. Results

4.1 Text Properties

Table 2 shows the descriptive statistics with respect to the different metrics, ordered by effect size (Cohen's *d*), which allows a comparison of the *Wablieft* corpus with *De Standaard 2010*. Figure 1 shows the frequency distributions for the six best-scoring features. The distributions for the other features can be found in appendix A. The smaller the overlap between the distributions for the two corpora, the more discriminative a feature is.

As shown in Table 2 and confirmed by Figure 1, the metric that differentiates best between the two corpora is the average number of words per sentence with an effect size of 0.81, which is considered a strong effect. This is in agreement with the Wablieft guideline of writing short sentences. On average, sentences in Wablieft are 8.31 words long, compared to 14.19 words per sentence in De Standaard. Only the Leesindex A comes close, scoring an effect size of 0.79, just under what is generally considered the (soft) threshold of 0.80 for a strong effect size. Note that average sentence length is one of the two variables (together with number of syllables per word) included in the Leesindex A (see Equation 6).

The syntactic metrics *tree depth* and *clause length* also have a fairly high effect size (0.77 and 0.73 respectively). These are followed by the Flesch Douma and Flesch readability metrics.

The long word ratio is the lexical measure with the highest effect size (0.67), followed by average age of acquisition (0.63), characters per word (0.62) and the Guiraud index (0.61). Frequent word ratio, which can be seen as the operationalisation of the Wablieft guideline to use everyday vocabulary, almost achieves a medium effect size (0.48).

The articles in Wablieft are, on average, more than twice as short as those in De Standaard (163 *vs* 383 words), but the high standard deviations show that there is considerable variation in both corpora. This is also clearly visible in the corresponding graph in appendix A. The relatively small effect size for this parameter (0.42) indicates that text length alone is not sufficient to differentiate between both corpora.

At the lower end of the spectrum, the part-of-speech ratios all show very small effect sizes, and in all cases score lower than the non-normalised part of speech frequency, which are related to text length.

	Wablieft			De Standaard			Cohen's d
Metric	**MEAN**	**MED**	**STDEV**	**MEAN**	**MED**	**STDEV**	Effect size
w/sen	8.31	8.19	1.14	14.19	13.96	3.12	**0.810**
Leesindex A	72.84	73.28	9.13	50.20	50.28	11.29	0.791
depth	4.44	4.40	0.49	6.12	6.11	0.92	0.774
cllen	8.13	8.00	1.15	11.46	11.31	2.09	0.728
Flesch Douma	77.80	78.39	10.10	59.84	60.01	10.84	0.722
Flesch	65.13	65.77	11.10	45.43	45.62	11.90	0.721
subcl/cl	0.11	0.10	0.09	0.27	0.28	0.11	0.675
lwr	0.04	0.03	0.03	0.08	0.08	0.03	0.670
sublen	5.56	5.42	2.01	8.38	8.17	2.38	0.643
CILT	77.66	77.84	4.96	70.46	70.64	5.45	0.641
NPlen	3.31	3.25	0.51	4.29	4.20	0.79	.0.635
aoa	7.92	7.83	1.06	9.55	9.43	1.29	0.630
c/w	4.76	4.75	0.34	5.23	5.21	0.36	0.623
gi	7.29	6.94	1.39	10.12	10.01	2.45	0.610
syl/w	1.58	1.57	0.13	1.74	1.73	0.13	0.607
fwr	78.29	78.44	6.00	72.88	73.62	6.19	0.479
ART	16.34	13.00	13.20	41.09	35.00	33.53	0.468
PRP	20.80	15.00	17.49	52.84	44.00	43.23	0.467
w/finV	8.12	8.00	1.29	10.32	9.86	2.99	0.457
subcl	3.05	2.00	5.00	12.64	9.00	14.23	0.457
N	41.69	30.00	31.31	88.34	76.00	68.17	0.426
SPEC	6.73	4.00	8.04	20.35	15.00	21.03	0.425
w/txt	163.68	104.00	144.95	383.19	315.00	329.75	0.420
ADJ	10.86	8.00	11.25	28.69	22.00	27.42	0.419
V	26.88	18.00	27.40	61.93	49.00	58.67	0.379
CONJ	4.99	3.00	6.20	17.63	13.00	18.22	0.374
PUN	24.04	16.00	24.94	55.58	43.00	53.77	0.374
ADV	11.38	7.00	14.02	25.01	17.00	27.38	0.318
cl	21.97	14.00	21.91	42.35	33.00	40.93	0.311
finV	21.10	14.00	20.74	40.40	32.00	38.90	0.310
Nr	0.24	0.24	0.04	0.21	0.21	0.04	0.308
NUM	4.49	4.00	4.20	8.40	6.00	8.80	0.291
CLIB	63.49	63.61	4.44	66.26	66.73	6.43	0.270
PRN	19.24	11.00	26.25	38.46	24.00	48.62	0.256
PRNr	0.09	0.09	0.04	0.08	0.07	0.03	0.242
ttr	0.63	0.63	0.09	0.59	0.57	0.11	0.231
PRPr	0.11	0.11	0.03	0.12	0.12	0.02	0.230
ADJr	0.06	0.06	0.02	0.06	0.06	0.02	0.199
d/h	1.72	1.72	0.11	1.75	1.76	0.10	0.190
SPECr	0.04	0.03	0.04	0.05	0.04	0.04	0.168
NUMr	0.03	0.02	0.02	0.02	0.02	0.02	0.156
ld	0.52	0.53	0.06	0.51	0.51	0.05	0.151
conc	2.93	2.93	0.24	2.96	2.96	0.18	0.116
ADVr	0.06	0.05	0.03	0.05	0.05	0.02	0.112
ARTr	0.09	0.09	0.03	0.10	0.10	0.03	0.098
Vr	0.14	0.14	0.03	0.14	0.14	0.03	0.061
CONJr	0.02	0.02	0.01	0.04	0.04	0.01	0.020
PUNr	0.12	0.12	0.03	0.12	0.12	0.03	0.020

Table 2: Statistical information concerning the different metrics. A Cohen's d-value > 0.80 (**indicated in boldface**) points to a strong effect. Values > 0.50 signify a medium effect. Values > 0.20 signify a small effect. These groups of effect sizes are delimited with a dashed line.

4.2 Correlations between features

Many of the linguistic features and readability metrics included in this study are logically and/or mathematically related. In this section we therefore explore the correlations between the different measures. We identified three clusters of inter-correlating features:[16]

1. Features related to text length: text length itself (w/txt), morpho-syntactic metrics gauging the occurrence of word categories (PRP, PRN, N, ART, CONJ, ADV, PUN and V), counts of syntactic units (cl, subcl and finV), and lexical diversity measures containing text length as denominator (gi and ttr). The correlation matrix of this cluster is presented in Table 3. Text length correlates very strongly ($r > .90$) with 10 of these features. It is not surprising, for example, that the number of specific tokens in a text is, to a large degree, a function of the total number of tokens in a text, or that longer texts contain more (sub)clauses. Also, the negative correlation between ttr and text length confirms that ttr should not be used as a measure of lexical diversity with texts of varying lengths, since longer texts tend to receive lower ttr scores (as it becomes more difficult to introduce new words in a longer texts).

	cl	**subcl**	**finV**	**ttr**	**gi**	**PRP**	**PRN**	**N**	**ART**	**CONJ**	**ADV**	**ADJ**	**PUN**	**V**
w/txt	.98	.85	.98	-.74	.84	.96	.94	.96	.90	.88	.93	.92	.96	.98
cl		.89	1.00	-.71	.82	.92	.96	.92	.85	.89	.94	.91	.97	.99
subcl			.88	-.59	.67	.75	.88	.74	.67	.87	.84	.80	.82	.88
finV				-.71	.82	.92	.96	.92	.85	.88	.94	.91	.97	.99
ttr					-.58	-.74	-.64	-.75	-.74	-.64	-.66	-.66	-.68	-.70
gi						.81	.76	.83	.75	.75	.76	.79	.82	.81
PRP							.86	.95	.89	.80	.84	.85	.90	.91
PRN								.83	.74	.87	.93	.89	.94	.96
N									.92	.82	.83	.85	.91	.90
ART										.72	.76	.80	.83	.84
CONJ											.85	.84	.86	.89
ADV												.88	.91	.94
ADJ													.89	.91
PUN														.96

Table 3: Highly correlating features related to text length.

2. Features related to sentence length: w/sen, cllen and (somewhat less) w/finV. Correlations are presented in Table 4. Sentences can be made longer by increasing the length of the (finite) clauses that they contain, or by combining several clauses into the same sentence.[17]

	cllen	**w/finV**
w/sen	.79	.54
cllen		.72

Table 4: Correlating features related to sentence length.

3. Readability metrics (Flesch, Flesch Douma, CILT, Leesindex A), and their constituting sub-features (char/w and syl/w). To a somewhat lesser extent, CLIB is also correlated with these, as presented in Table 5. We see a perfect[18] correlation between Flesch and Flesch Douma, and a nearly perfect correlation between both of these measures and Leesindex A.

16. All correlation coefficients presented in this section are based on the Wablieft data set.
17. Note that the Wablieft corpus contains mainly short, simple sentences.
18. Correlation coefficient $r > .999999$

	Flesch Douma	CLIB	CILT	Leesindex A	char/w	syl/w
Flesch	1.00	.83	.55	.99	-.87	-.99
	Flesch Douma	.83	.55	.99	-.87	-.99
		CLIB	.86	.48	-.69	-.60
			CILT	.80	-.95	-.84
				Leesindex A	-.84	-.97
					char/w	.88

Table 5: Correlating readability metrics and their highly correlating constituting features.

We decided to take the high inter-correlations between measures into account for the automated text classification using binary classifiers, which is presented in the next section. For the models using multiple features, we worked with a reduced feature set by eliminating certain measures that were found to correlate very strongly, in order to reduce the search space for the A* algorithm and to avoid multi-collinearity. For the measures related to text length, we dropped the features cl, finV, ADV, PUN, V, PRN, N, ADJ, SPEC, PRP, NUM and ART. We also removed Flesch, Flesch Douma and syl/w from the cluster of readability metrics.

4.3 Classification results

When we look at the two different classification methods we tried (described in section 3.4), the accuracy of the neural networks was always lower than that of the corresponding logistic regression models. This was the case for single features as well as for combinations of features. We tried different network sizes and architectures, and the best score we reached for a single feature (w/sen) in a neural classifier was 91.41% (compared to 94.46% for the corresponding logistic regression model). The best neural score for all features combined was 97.58% (compared to 98.54% for the best logistic regression model). The logistic regression models score very high overall and are, moreover, much faster to train than the neural nets. Since we failed to show an added value of using neural networks for this particular task and data set, we only present detailed results for the logistic regression models.

Features	Score	Features	Score	Features	Score
w/sen	0.9446	ART	0.7563	NUM	0.6413
depth	0.9042	w/txt	0.7538	CLIB	0.6292
cllen	0.8763	subcl	0.7508	Nr	0.6188
Leesindex A	0.8750	w/finV	0.7475	ttr	0.6183
Flesch	0.8129	syl/w	0.7450	PRPr	0.5779
Flesch Douma	0.8117	V	0.7292	PRNr	0.5758
NPlen	0.8067	N	0.7292	ADJr	0.5742
sublen	0.8063	ADJ	0.7254	d/h	0.5692
lwr	0.7917	PUN	0.7254	SPECr	0.5579
subcl/cl	0.7913	CONJr	0.7021	conc	0.5567
aoa	0.7838	SPEC	0.6983	NUMr	0.5554
gi	0.7738	cl	0.6929	ld	0.5475
CILT	0.7696	finV	0.6917	ARTr	0.5433
c/w	0.7633	fwr	0.6779	ADVr	0.5417
CONJ	0.7617	ADV	0.6650	PUNr	0.5283
PRP	0.7571	PRN	0.6438	Vr	0.5046

Table 6: Results of logistic regression with one feature only.

If we look at single features only (Table 6), the best scoring measure is w/sen, as was the case when looking at effect size. Not surprisingly, the rankings of the features according to their effect size and their classification accuracy are extremely well correlated (Spearman's $\rho = 0.97$).

With regard to the traditional readability metrics, Leesindex A, Flesch Douma and Flesch score above 80% in the logistic regression. In fact, these three metrics contain the same two parameters, namely words per sentence and syllables per words, and highly correlate ($r > .98$). When these two features are entered as independent variables in a logistic regression, we obtain a classification accuracy of 95.17%, which is well above the results we got when using the individual readability metrics as independent variable. It seems that, for our data at least, a better combination of words per sentence and syllables per word is possible than the ones used in the pre-defined readability metrics.

Concerning CLIB, we see that this measure only scores 62.92% in the logistic regression, which is surprisingly low (and the lowest of all readability metrics we implemented). Each of the four constituting features, apart from ttr, scores better by itself. In Table 7 we present the scores for different combinations of the four features that are used to calculate CLIB. When using only words per sentence and characters per word, we obtain an accuracy of 95.63%. Note that characters per word and syllables per word, which are used in Leesindex A, Flesch Douma and Flesch, are closely related features ($r = .88$). Adding ttr only slightly increases the accuracy.

2 features		**3 features**		**4 features**	
w/sen, c/w	0.9563	w/sen, c/w, ttr	0.9575	all	0.9550
w/sen, fwr	0.9475	w/sen, c/w, fwr	0.9554		
w/sen, ttr	0.9429	w/s, fwr, ttr	0.9483		
c/w, fwr*	0.7721	char/w, fwr, ttr	0.8200		
c/w, ttr	0.7842				
fwr, ttr	0.7488				

Table 7: Results of logistic regression for the combination of features involved in CILT (marked with *) and CLIB.

The logistic regression model using CILT performs better, obtaining an accuracy of 76.96%. CILT offers a linear combination of two features, namely frequent word ratio and characters per word. Both of these features score, by themselves, slightly lower than CILT. The combination of the features, as shown in Table 7, reaches an accuracy of 77.21%, which is only slightly better than CILT. Nevertheless, CILT is not using the best predictors for our classification task.

Next, we look at models including combinations of multiple features, based on the A* search described in section 3.4 and disregarding certain highly correlated measures as identified in section 4.2. The best-scoring feature combinations up to seven features are listed in Table 8, which also included the average accuracies of the neural classifier using the same features by means of comparison. When using more than seven features, several different combinations lead to the same best scores. Table 8 shows that by combining the syntactic metric w/sen with different lexical metrics (such as aoa, gi, lwr and ttr), a readability formula (CLIB or CILT), a morpho-syntactic metric (Nr) and, to a lesser extent, other syntactic metrics (depth, cllen), the accuracy of the classification can be further improved.

# features	Accuracy		Features
	logit	MLP	
1	0.9446	0.9222	w/sen
2	0.9575	0.9486	w/sen + aoa
3	0.9721	0.9661	w/sen + CLIB + gi
4	0.9729	0.9518	w/sen + CLIB + gi + lwr
5	0.9758	0.9673	w/sen + CLIB + gi + Nr + ttr
6	0.9771	0.9677	w/sen + CLIB + gi + Nr + ttr + depth
7	0.9788	0.9740	w/sen + CILT + gi + Nr + ttr + cllen + aoa

Table 8: Best feature combinations in logistic regression (logit). Average scores (over 30 runs) for the multilayer perceptron classifier using the same features are also mentioned.

4.4 Ignoring sentence length (w/sen)

Considering that the predictive value of sentence length alone is so strong (for our data set specifically, but also according to previous analyses of text readability), we wanted to conduct an additional experiment in which we control for differences in w/sen across articles. To this end, we extracted from both subcorpora all articles with w/sen values between 6 and 12, which are the values for which we see an overlap in w/sen in Figure 1. Table 9 shows how many articles in each subcorpus occur with certain w/sen values. Each of the columns represents a bin.[19] In order to allow recalculating the effect of each of the features (without w/sen), for each bin we determine the size of each subcorpus $N = min(size(Wablieft), size(DeStandaard))$, selecting N random articles from the larger corpus, so the two subcorpora are of equal size with respect to number of articles, allowing easy comparison.

	$6 \leq x < 7$	$7 \leq x < 8$	$8 \leq x < 9$	$9 \leq x < 10$	$10 \leq x < 11$	$11 \leq x < 12$
Wablieft	1160	3994	4262	2075	800	218
De Standaard	140	206	446	990	1918	3192
N	140	206	446	990	800	218

Table 9: Number of newspaper articles for specific ranges of w/sen values.

Table 10 shows the best scoring single features in the logistic regression when controlling for w/sen and Figure 2 shows the best scores for logistic regression classifiers found using the same A* algorithm as described in section 4.3, but now over bins with comparable w/sen values. The values in this table and figure are weighted averages of the values of the bins as calculated according to Equation 7, in which n is the number of the bin, N_n the size of the bin (as in Table 9), and μ_n the mean value calculated for that specific bin.

	$6 \leq x < 7$	$7 \leq x < 8$	$8 \leq x < 9$	$9 \leq x < 10$	$10 \leq x < 11$	$11 \leq x < 12$	AVG
CILT	0.8571	0.7619	0.8111	0.7576	0.7125	0.8409	0.7650
CLIB	0.8214	0.7381	0.8000	0.7222	0.7125	0.7273	0.7384
aoa	0.8929	0.7619	0.7889	0.6818	0.6750	0.7727	0.7204
fwr	0.8214	0.6905	0.7667	0.6869	0.7250	0.6818	0.7171
lwr	0.7857	0.7857	0.7333	0.7525	0.6125	0.7273	0.7116

Table 10: Best scoring features when controlling for w/sen.

19. In the remainder of this section, we will shorten the names of the bins to bin6, bin7, bin8, bin9, bin10 and bin11 respectively.

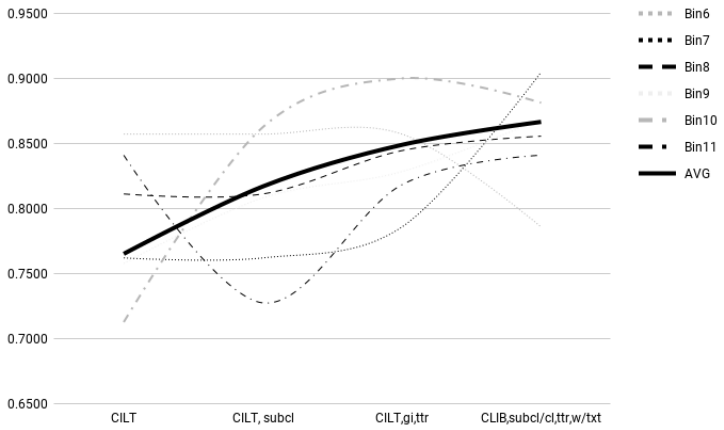

Figure 2: Weighted average of best logistic regression results for each of the bins of equal sentence length over number of features.

$$AVG(X) = \frac{\sum_{n=6}^{11} N_n * \mu_n}{\sum_{n=6}^{11} N_n} \tag{7}$$

When ignoring w/sen and only using one other feature, CILT is the best predictor (admittedly it is a combination of two features). With two features, CILT combined with subcl scores best, and with three features the best combination includes CILT, gi and ttr.[20] With four features, we find that CLIB (which correlates very strongly with CILT) combined with subcl/cl, ttr and w/txt shows the best prediction.

The weighed average clearly shows diminishing improvement when adding features, but we still reach an average accuracy of over 85%.

5. Conclusions and Future Work

We first present our conclusions in Section 5.1. In Section 5.2 we present our plans for the future.

5.1 Conclusions

In this paper we identified a number of linguistic features or metrics that make good proxies for readability by using the Wablieft corpus as a gold standard of easy-to-read texts, and the De Standaard data as a control corpus. The simple metric words/sentence scores best of all, both in terms of effect size and classification accuracy. Generally speaking, syntactic metrics appear to perform better than lexical ones, even though adding lexical metrics (such as average age of acquisition, Guiraud or characters/word) to average sentence length increases the accuracy of classification using logistic regression. We also found that of the traditional metrics, which all combine syntactic and lexical features (apart from CILT), Brouwer's Leesindex A scores best. It should be noted, however, that none of these measures achieve the effect size or accuracy of words per sentence alone, which indicates that, for our data at least, there is no evidence supporting their use as readability metrics.[21] For the classification task we used here, better results could be reached in the logistic regression when using features as separate independent variables. Nevertheless, additional analyses

20. Note that gi is actually a transformation of ttr.
21. All these metrics, apart from CILT include w/sen as a factor.

showed that, when controlling for differences in average sentence length, the CILT metric proved to have the best predictive power.

We found three clusters of inter-correlated features, i.e. features related to text length, features related to sentence length, and the readability metrics. We argue that such (high) inter-correlations between measures can (and maybe should) be taken into account in future classification tasks.

We also noted that the details of operationalisation of the metrics can make a huge difference for the results of our analyses. For instance, how you define content words or how you count syllables, what you do with words that do not occur in the age-of-acquisition or concreteness data sets has its effect on the evaluation of the usefulness of the metrics with respect to readability. Also the corpus that is used to determine word frequency plays a big role.

Furthermore, details on how you filter texts and treebanks to distinguish noise from language can make a considerable difference. Our ranking of best proxies looked quite different before we removed about 1000 articles from De Standaard while cleaning the data.

It is important to keep in mind that our operationalisation of readability was based on a comparison of easy-to-read newspaper articles produced by trained and professional writers of such text with regular newspaper articles, rather than on actual judgements of (perceived) readability. This, of course, has both advantages and disadvantages. It is probably an advantage to evaluate the readability of texts based on actual texts written by expert writers who follow clear norms for writing *simple* texts. Yet, since many of the instructions concerning the writing of easy-to-read texts make reference to certain features that we evaluated in this study (such as sentence length, lexical variety and frequency), there is certainly a risk of circular reasoning in the evaluations we carried out. Moreover, considering that using a simple metric such as average sentence length already enabled us to reach 94.5% classification accuracy, it might be argued that our two corpora were actually too different (in terms of readability-related aspects). Since our simple 1-feature classifier already worked so well, there was not much room for building better classification models using a wider, possibly more varied range of features. Because of the design of this study, we also only relied on a binary distinction between easy-to-read and regular text, which does not constitute the most refined assessment of readability.

5.2 Future work

In future work, we want to identify comparable articles (i.e. dealing with the same topic) from De Standaard en Wablieft, and build a comparable corpus with on the one side *regular* Dutch texts and on the other side *easy* Dutch texts.

From that corpus, we want to identify comparable sentences, roughly treating the same content. As such, we hope to build a corpus of comparable sentences with on the one side *regular* Dutch sentences and on the other side *easy* Dutch sentences. As regular sentences are often longer, one-to-many mappings will regularly occur.

Based on these data sets we aim at building an automated text simplifier using supervised (Wubben et al. 2012), unsupervised or semi-supervised MT technology, for example using neural MT systems (Nisioi et al. 2017, Wang et al. 2016), as alternative to rule-based methods for lexical (Bulté et al. 2018) and syntactic simplification (Sevens et al. 2018) in Dutch or to data-driven tree transduction methods for Dutch sentence compression (Vandeghinste and Pan 2004).

References

Aluísio, Sandra M., Lucia Specia, Thiago A. S. Pardo, Erick G. Maziero, Helena M. Caseli, and Renata P. M. Fortes (2008), A corpus analysis of simple account texts and the proposal of simplification strategies: First steps towards text simplification systems, *Proceedings of the 26th Annual ACM International Conference on Design of Communication*, SIGDOC '08, ACM, New York, NY, USA, pp. 15–22. http://doi.acm.org/10.1145/1456536.1456540.

Augustinus, Liesbeth, Vincent Vandeghinste, Ineke Schuurman, and Frank Van Eynde (2017), GrETEL: A Tool for Example-based Treebank Mining, *CLARIN in the Low Countries*, London: Ubiquity Press, chapter 22, pp. 269–280.

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

Brysbaert, M., M. Stevens, S. De Deyne, W. Voorspoels, and G. Storms (2014), Norms of age of acquisition and concreteness for 30,000 Dutch words, *Acta Psychologica* **150**, pp. 80–84.

Bulté, Bram, Leen Sevens, and Vincent Vandeghinste (2018), Automating lexical simplification in Dutch, *Computational Linguistics in the Netherlands Journal* **8**, pp. 24–48. https://clinjournal.org/clinj/article/view/78.

Cardon, Rémi and Natalia Grabar (2018), Identification of parallel sentences in comparable monolingual corpora from different registers, *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, Association for Computational Linguistics, pp. 83–93. http://aclweb.org/anthology/W18-5610.

Collins-Thompson, Kevyn and James P. Callan (2004), A language modeling approach to predicting reading difficulty, *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, Association for Computational Linguistics, Boston, Massachusetts, USA, pp. 193–200. https://www.aclweb.org/anthology/N04-1025.

Cucchiarini, Catia, Joris Driesen, Hugo Van hamme, and Eric Sanders (2008), Recording speech of children, non-natives and elderly people for HLT applications: the JASMIN-CGN corpus., *LREC 2008*. http://www.lrec-conf.org/proceedings/lrec2008/pdf/366_paper.pdf.

De Clercq, Orphée and Véronique Hoste (2016), All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch, *Computational Linguistics* **42** (3), pp. 457–490. https://www.aclweb.org/anthology/J16-3004.

De Clercq, Orphée, Sarah Schulz, Bart Desmet, Els Lefever, and Véronique Hoste (2013), Normalization of Dutch user-generated content, *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria, pp. 179–188. https://www.aclweb.org/anthology/R13-1024.

De Clercq, Orphée, Véronique Hoste, Bart Desmet, Philip Van Oosten, Martine De Cock, and Lieve Macken (2014), Using the crowd for readability prediction, *Natural Language Engineering* **20** (3), pp. 293–325, Cambridge University Press.

de Medeiros Caseli, Helena, Tiago de Freitas Pereira, Lucia Specia, Thiago A. S. Pardo, Caroline Gasperin, and Sandra M. Aluísio (2009), Building a Brazilian Portuguese parallel corpus of original and simplified texts, *Proceedings of CICLing*.

Dechter, Rina and Judea Pearl (1985), Generalized Best-first Search Strategies and the Optimality of A*, *Journal of ACM* **32** (3), pp. 505–536, ACM, New York, NY, USA. http://doi.acm.org/10.1145/3828.3830.

Dell'Orletta, Felice, Simonetta Montemagni, and Giulia Venturi (2011), READ–IT: Assessing readability of Italian texts with a view to text simplification, *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, Association for Computational Linguistics, Edinburgh, Scotland, UK, pp. 73–83. https://www.aclweb.org/anthology/W11-2308.

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

Feng, Lijun, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad (2010), A comparison of features for automatic readability assessment, *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 276–284.

Flesch, R. (1948), A new readability yardstick, *Journal of Applied Psychology* **32**, pp. 221–233.

François, Thomas (2009), Combining a statistical language model with logistic regression to predict the lexical and syntactic difficulty of texts for FFL, *Proceedings of the Student Research Workshop at EACL 2009*, Association for Computational Linguistics, Athens, Greece, pp. 19–27. https://www.aclweb.org/anthology/E09-3003.

Gonzalez-Dios, Itziar, María Jesús Aranzabe, and Arantza Díaz de Ilarraza (2018), The corpus of Basque simplified texts (CBST), *Language Resources and Evaluation* **52** (1), pp. 217–247. https://doi.org/10.1007/s10579-017-9407-6.

Grabar, Natalia and Rémi Cardon (2018), CLEAR – Simple corpus for medical French, *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, Association for Computational Linguistics, pp. 3–9. http://aclweb.org/anthology/W18-7002.

Guiraud, P. (1954), *Les Caractres Statistiques du Vocabulaire. Essai de mthodologie*, Presses Universitaires de France, Paris.

Hancke, Julia, Sowmya Vajjala, and Detmar Meurers (2012), Readability classification for German using lexical, syntactic, and morphological features, *Proceedings of COLING 2012*, The COLING 2012 Organizing Committee, Mumbai, India, pp. 1063–1080. https://www.aclweb.org/anthology/C12-1065.

Inclusion-Europe (2009), Information for all. European standards for making information easy to read and understand. https://easy-to-read.eu/wp-content/uploads/2014/12/EN_Information_for_all.pdf.

Larsson, Patrik (2006), *Classification into Readability Levels. Implementation and Evaluation*, PhD thesis, Uppsala University, Sweden.

MacWhinney, Brian (2000), *The CHILDES project: The database*, Vol. 2, Psychology Press.

Mühlenbock, Katarina (2009), Readable, legible or plain words - Presentation of an easy-to-read Swedish corpus, *Multilingualism, Proceedings of the 23rd Scandinavian Conference of Linguistics*, Vol. Studia Linguistica Upsaliensia 8, pp. 325–327.

Naderi, Babak, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller (2019), Subjective assessment of text complexity: A dataset for German language, *CoRR*. http://arxiv.org/abs/1904.07733.

Nisioi, Sergiu, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu (2017), Exploring neural text simplification models, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 85–91.

Odijk, Jan, Gertjan van Noord, Peter Kleiweg, and Erik Tjong Kim Sang (2017), The parse and query (PaQu) application, *CLARIN in the Low Countries*, London: Ubiquity Press, chapter 23, pp. 281–297.

Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste, and Ineke Schuurman (2013), The construction of a 500-million-word reference corpus of contemporary written Dutch, *in* Spyns, Peter and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch. Results by the STEVIN programme.*, Springer.

Pander Maat, Henk, Rogier Kraf, Antal van den Bosch, Nick Dekker, Maarten van Gompel, Suzanne Kleijn, Ted Sanders, and Ko van der Sloot (2014), T-scan: a new tool for analyzing Dutch text, *Computational Linguistics in the Netherlands Journal* **4**, pp. 53–74. https://clinjournal.org/clinj/article/view/40.

Schwarm, Sarah and Mari Ostendorf (2005), Reading level assessment using support vector machines and statistical language models, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, Association for Computational Linguistics, Ann Arbor, Michigan, pp. 523–530. https://www.aclweb.org/anthology/P05-1065.

Sevens, Leen, Vincent Vandeghinste, Ineke Schuurman, and Frank Van Eynde (2018), Less is more: A rule-based syntactic simplification module for improved text-to-pictograph translation, *Data and Knowledge Engineering*.

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

Tack, Anaïs, Thomas François, Piet Desmet, and Cédrick Fairon (2018), NT2Lex: A CEFR-graded lexical resource for Dutch as a foreign language linked to open Dutch WordNet, *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, Association for Computational Linguistics, New Orleans, Louisiana, pp. 137–146. https://www.aclweb.org/anthology/W18-0514.

United Nations (1984), *Handbook of Household Surveys, Revised Edition*, Vol. 31 of *Studies in Methods, Series F*, United Nations, New York.

Van Eynde, Frank (2004), Part of speech tagging and lemmatizing of the Corpus Gesproken Nederlands (Spoken Dutch Corpus). http://nederbooms.ccl.kuleuven.be/documentation/manual-EN-POS-CGN.pdf.

van Noord, Gertjan (2006), At last parsing is now operational, *TALN 2006*, pp. 20–42.

Vandeghinste, Vincent and Yi Pan (2004), Sentence compression for automated subtitling: A hybrid approach, *Text Summarization Branches Out*, Association for Computational Linguistics, Barcelona, Spain, pp. 89–95. https://www.aclweb.org/anthology/W04-1015.

Vandeghinste, Vincent, Bram Bulté, and Liesbeth Augustinus (2019), Wablieft: An easy-to-read newspaper corpus for Dutch, *Proceedings of the CLARIN Annual Conference*, Leipzig, Germany, pp. 188–191.

Wang, Tong, Ping Chen, John Rochford, and Jipeng Qiang (2016), Text simplification using neural machine translation, *AAAI Conference on Artificial Intelligence*, pp. 4270–4271.

Wubben, Sander, Antal van den Bosch, and Emiel Krahmer (2012), Sentence simplification by monolingual machine translation, *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Jeju Island, Korea, pp. 1015–1024. https://www.aclweb.org/anthology/P12-1107.

Yaneva, Victoria (2015), Easy-read documents as a gold standard or evaluation of text simplification output, *Proceedings of the Student Research Workshop*, INCOMA Ltd. Shoumen, Bulgaria, pp. 30–36. http://aclweb.org/anthology/R15-2005.

Appendix A. Distribution of features

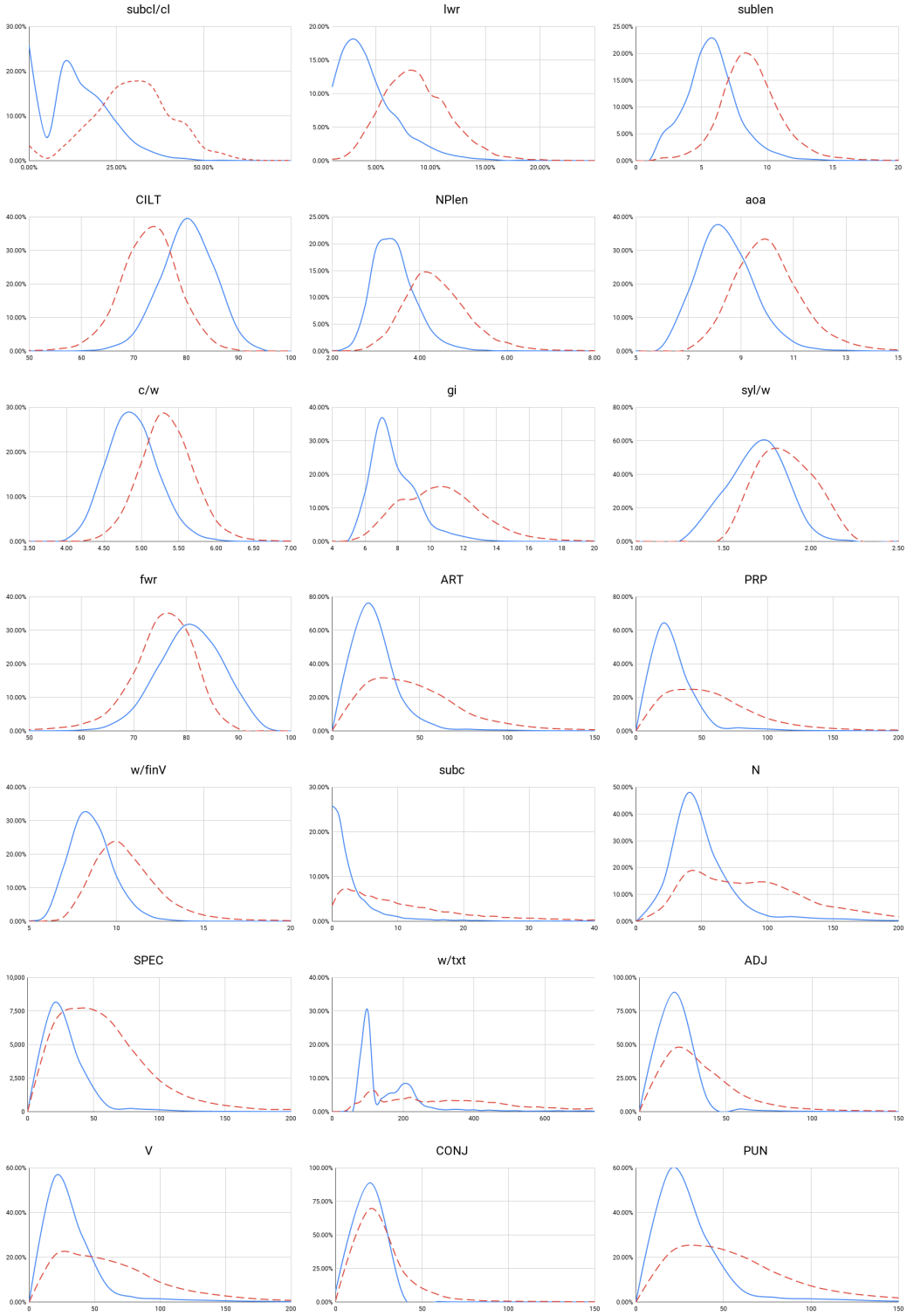

Figure 3: Continuation of Figure 1.

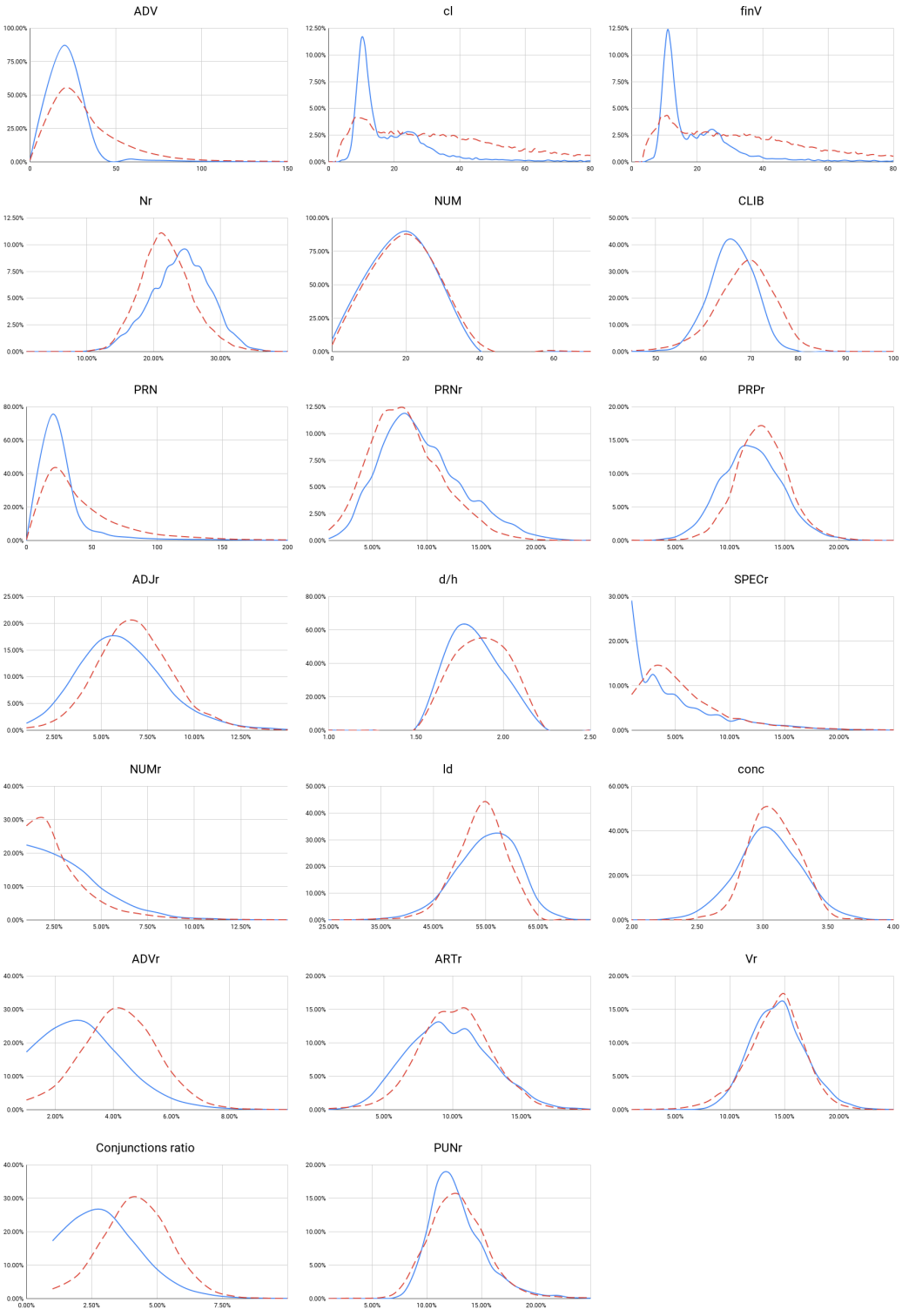

Figure 3: (continued)