Natural Language Processing with Python: Term Paper

Applying Readability Formulae to the Task of Proficiency Assessment

Zarah Leonie Weiß*

This paper investigates the performance of readability formulae for German second language (L2) proficiency assessment. The change of application domain shows quite promising results with f1-scores up to 51.5% for the Coleman Liau Index on the German subsection of the Merlin corpus. It also introduces a freely available program for calculating readability formula for German texts, the Readability Calculator, which is freely available for use at `https://github.com/zweiss/RC_Readability_Calculator`.

Keywords: Proficiency assessment, readability formulae, Merlin corpus

1. Introduction

The reliable and robust assessment of text readability with computational methods is an important task for all areas in which texts are produced or selected. This holds especially for first and second language learning contexts, on which a large amount of past and present research has focused (cf. Crossley, Greenfield, and McNamara (2008), DuBay (2004), Greenfield (2003)). Yet, readability assessment is also highly interesting from an economical perspective for companies selling texts, such as news papers or books. They have an interest in producing texts appealing for a broad audience, hence there is a need of controlling for the appropriate reading ease of their texts. This holds for news paper articles (DuBay 2004; McLaughlin 1969), but also for business to business communication (Leong, Ewing, and Pitt 2002) and many related fields. Aside from the educational and ecological perspective, text readability is also often discussed in the context of technical or medical instructions or political texts, thus being discussed in the context of inclusive distribution of information relevant for everyone no matter their reading skills see e.g. (Esfahani et al. 2014; Janghorban Esfahani et al. 2015; von der Brück 2007, 2008). Accordingly, the value of reliable and accessible measures of text readability can hardly be overestimated.

Often, readability formulae are considered to serve as such reliable and accessible measures of readability. They provide readability indices, which are scaled to allow for convenient interpretations, based on a small set of often superficial linguistic features, such as length measures of linguistic units, or word list frequencies. Also, they are available for a broad range of languages, such as Spanish, French, German, Dutch, Swedish, Russian, Hebrew, Hindi, Chinese, Vietnamese, and Korean (cf. DuBay (2004, p. 57)), and there is a series of platforms offering to calculate various readability scores

* Wilhelmstraße 19, 72074 Tübingen, R. 1.29. E-mail: zarah-leonie.weiss@uni-tuebingen.de.

with or without charging fees. Hence, their frequent use in many research contexts is hardly surprising. Despite their popularity readability formulae are controversial, though. In fact, they are often criticised for being too simplistic in terms of scaling and feature selection. Notwithstanding this criticism, which is addressed in broader detail in section 3, readability formluae have been shown to correlate highly with reading measurements in experimental settings (cf. (DuBay 2004)). Also, superficial features such as sentence or word length are highly correlated with more fine-grained features associated with linguistic complexity, which again are related to readability (Greenfield 2004, p. 7).

If language complexity is indirectly measured by readability formulae, they might not only be suited to estimate readability, but also L2 proficiency, as this also is often assessed in terms of linguistic complexity (Vajjala and Meurers 2012; Hancke, Vajjala, and Meurers 2012; François and Fairon 2012). Since learner writing is highly non-regular and a challenge to Natural Language Processing (NLP) tools, it might be beneficial to identify the informativeness of superficial features and readability formulae. Therefore, this paper investigates the performance of readability formulae on the task of proficiency assessment of L2 German, showing encouraging results: all readability formulae perform significantly above chance and the Coleman Liau Index (CLI) achieves a f1-score of 51.5%. For formula calculation a new tool to obtain various readability scores for German based on plain text input was written using Python and the Natural Language Tool Kit (NLTK)[1]: the Readability Calculator (RC) is freely available at `https://github.com/zweiss/RC_Readability_Calculator`. It may not only be used to obtain readability scores. Unlike most platforms it also allows insights in the concrete implementation details, making the obtained scores more transparent, see section 5 for more details.

The remainder of this paper is structured as follows: First, related work on readability formulae is discussed. This is followed by a brief section providing some background information on benefits and downsides of readability formulae. Then, the Merlin corpus is introduced, which is used for both, the evaluation of the RC as well as the experiments. Section 5 elaborates on the RC, discussing the processing pipeline, the readability formulae implemented, and an evaluation of the tool's feature calculation reliability measured in terms of precision, recall and f1-score on a random subset of the Merlin corpus. Afterwards, the classification experiments are presented and discussed in section 6. The paper closes with a note on possible future work and a conclusion.

2. Related Work

The earliest readability formula was designed by Lively and Pressey (1923) (cf. DuBay (2004, p. 14)) and was used to access the readability of school books. In the last decade readability formulae have been used especially broadly in the assessment of web page readability: Leong, Ewing, and Pitt (2002) use Dale-Chall grade level, Flesch Reading Ease (FRE), and Fog index to assess the readability of business-to-business web pages. Janghorban Esfahani et al. (2015), Esfahani et al. (2014), Luers et al. (2013) evaluate the readability of medical information from various domains provided on web pages of German medical centres to inform interested non-professionals. They use five readability formulae: Amstad Lesbarkeits Index (ALI), German SMOG index, Wiener Sachtext Formel (WSF), and Hohenheimer Verständlichkeitsindex.

1 `http://www.nltk.org/`.

However, readability formulae are also employed on other text domains. The University of Hohenheim regularly issues press releases concerning the readability of texts of political or economical significance, such as the FAQ sections offered by over 100 companies (Brettschneider 2012), the customer communication provided by telecommunication companies (Brettschneider 2013), CEO speeches at the general meeting of the "DAX 30" companies (Brettschneider 2014), and election programs in Baden-Württemberg state from 1980 to 2016 (Brettschneider 2016). They use their own readability formula, the Hohenheimer Lesbarkeitsindex. For a comprehensive overview of first language (L1) application domains for L1 readability assessment, please see DuBay (2004, p. 55).

While most readability formulae are employed in L1 contexts, there is also a tradition of accessing L2 readability with them. For this, either L1 or L2 formulae are used, the latter being specifically designed and parametrized for L2 contexts. Examples are the the English as a Foreign Language (EFL) Difficulty Estimate by Brown (1997) and the Miyazaki EFL Readability Index by Greenfield (1999), which were both designed for Japanese EFL speakers. There have been a lot of investigations on whether L1 and L2 formulae differ in applicability for L2 reading contexts, which lead to diverging conclusions: Hamsik (1984) reports a high positive correlation of L1 formula scores with cloze scores obtained from L2 English speakers. Greenfield (1999) reports similar results for Japanese EFL speakers, yet also finds that his L2 formula results in a slightly but significantly increased correlation. Yet, Brown (1992) does not obtain a sufficient correlation between L1 formula scores and Japanese EFL speaker performance. He concludes from the significantly higher correlation of his EFL Difficulty Estimate, that L2 formulae are in fact more suited for L2 readability assessment. However, it should be noted, that his formula does not exceed a R^2 of 0.51 either. Greenfield (2004) suggests that Brown (1992)'s results might also be explained by the broader range of textual domains used in the experiment, arguing that the L1 formulae used were parametrized for academic texts only. Greenfield (2004) employs both, L1 formulae and the EFL Difficulty Estimate on academic texts, finding high correlations with EFL reader performance for all formulae and no significant difference between L1 and L2 parametrization.

In general, L1 readability formulae are still broadly used for L2 readability assessment tasks as well (Crossley, Greenfield, and McNamara 2008) and there is a wider range of L1 readability formulae available for varying target languages than for L2 readability formulae, which makes the application of L2 formulae in practise more rare.

3. Readabilite Formulae

Readability formulae employ a very limited set of superficial, parametrized language features. They mainly rely on previously compiled word lists for vocabulary analyses and length measures of various linguistic units, such as sentences and words, for vocabulary and syntactic analyses. In order to obtain a readability formula from these superficial features, some initial theoretical considerations as to whether to a) use addition, multiplication or subtraction to combine them, and b) whether to transform them, for example as log transforms or by squaring are made. After this initial set up of a formula, a regression analysis is performed based on some training corpus, such as student's reading material, in order to identify the proper weights for each feature in the readability formula. Thus, readability formulae may be understood as regression equations (cf. Greenfield (2004, p. 6)). This procedure of turning a few linguistic features to simple regression equations makes readability formulae relatively easy to compute. They rely only on a very limited number of NLP tools, such as sentence segmentizers,

tokenizers, and sometimes syllable counters. In fact, many prominent readability formulae have been designed to work with a minimal amount of linguistic information, due to the technical limitations at the time of their development, for example the CLI by Coleman and Liau (1975).

While this simplicity in terms of feature selection and composition is often considered a great merit of readability formulae, it also leads to some of their major criticism towards that readability formulae fail to measure text complexity in sufficient detail or to take psycholinguistic insights into account (cf. DuBay (2004, p. 3), Greenfield (2004, p. 7), Bailin and Grafstein (2001, p. 290ff). Bailin and Grafstein (2001, p. 287ff)) especially question the use of vocabulary lists. They argue, that vocabulary use is highly dependent on the interaction between topic, genre, dialect, and sociolect of the text with the concrete reading group. Hence, they question whether it is appropriately generalizable to various reading groups. Despite these arguments, various studies have found them to correlate highly with other measures of readability. For example, Greenfield (2004, p. 7f) reports them to usually correlate with independent comrehension tests around 0.8 and 0.9. Hence, the validity of readability formulae is often argued to be externally verified, see also (DuBay 2004). Furthermore, the correlation of readability formulae with comprehension measures was found to either not increase significantly or to even decrease, when increasing the amount of features incorporated (DuBay (2004, p. 18), Greenfield (2004, p. 7)). Instead, pre-compiled vocabulary lists and length measures of linguistic units have proven to be most predictive in a variety of experiments (cf. DuBay (2004, p. 19)).

However, readability formulae are not suited to explain how readability of a text comes about, they merely predict it due to their high statistical correlation with a variety of elaborate linguistic features affecting readability (cf. Greenfield (2004, p. 8)): Examples are the correlation of sentence length with compounds and hypotactic sentence structures (Greenfield 2004, p. 7), but also the correlation of syllable counts with morpheme counts and Yngve's word depth (DuBay 2004, p. 19) and the known correlation between word frequency and word length (Zipf's law). Also, it should be noted, that the generalizability of readability formulae to varying reading scenarios is only partially investigated and in practise often ignored. Examples are the usage of L1 formulae in L2 contexts, the usage of formulae for English on other languages without re-parametrization, and the application of readability formulae on various text types. However, based on the data set used to parametrize and validate a readability formula, this formula is highly specific to what is referred to as the formula's *reading setting* for the remainder of the article: The reading setting consists of the reading group (e.g. students or adults), target language (e.g. English, German, French), reader L1, and text type (e.g. news paper articles, educational or web texts). While a formula may generalize well for certain aspects of this setting, it does not necessarily do so for others. For example, Greenfield (2004, p. 8, 15) argues, that readability formulae generalize across reader L1s, but that formulae designed for educational contexts cannot make any claims about readability for other text types. Please see Greenfield (2004, p. 8) and DuBay (2004) for further considerations regarding changes in reading group.

In practise readability formulae tend to be used with little concern towards their reading setting. This manifests in the usage of readability formulae parametrized for educational texts in studies on the readability of business texts (Leong, Ewing, and Pitt 2002) or medical texts (Esfahani et al. 2014; Janghorban Esfahani et al. 2015; Luers et al. 2013), but also in the lack of disclaimers towards the readability setting of formulae on the various platforms offering readability analyses. These platforms often provide easy access to formulae, but rarely offer all information needed to interpret them correctly.

The issue how to interpret readability scores is an important question for itself. On first glance, readability formulae seem to allow for an easy interpretation: They are either already scaled to some meaningful unit, such as the grade level for which a text is appropriate, or accompanied by a mapping of scales to ordinal estimates like relative adjectives (e.g. “easy" and “difficult"). These allow for a straight forward scale-immanent comparison of texts, i.e. given their scores, one may easily determine which text is more readable than the other. However, they are notoriously vague on the scale-external dimension: Even though some formulae come with a concrete reference to what the upper and lower end of their scales map to in terms of real world reading groups, the interpretation of the scores between the two poles remains notoriously difficult: What does it mean for a text to be “fairly easy"? This question relates to the general issue of vagueness of relative adjectives, which is well known and broadly discussed in the field of adjective semantics, see Kennedy and McNally (2005) for an introduction to the discussion. While sometimes attempted to be grounded with more or less concrete references, the transfer of such scores to concrete applications continues to be problematic, for example “fairly easy" corresponds to “slick fiction" and was readable for 83% of U.S. Americans in 1949 (DuBay 2004, p. 22). Grade scaling avoids this issue only to a certain extend. For example, a grade level scaled score returned by the Flesch Grade Scale (FGS) indicates that the average score in a comprehension test for an American school class of that grade level would be 75% (Greenfield 2003, p. 42). Yet, it is not quite clear how this relates to students who are not in the American school system, i.e. how readability formulae generalize on the synchronic axis. Also, the FGS was parametrized in 1978, hence, refers to average American school classes in the late 1970s. How this relates to contemporary school classes, i.e. how well formulae generalize on a diachronic axis, seems to be at least debatable.

Aside from the questions of interpretability and generalizability, the consistency of readability formula scales is another issue: Grade levels assigned to a text by various readability formulae have been shown to differ widely. For example, scores calculated by the Dale-Chall formula by Chall and Dale (1995) are known to be generally two grades lower than scores predicted by the SMOG formula by McLaughlin (1969) (cf. DuBay (2004, p. 47)). On a similar note, Leichter et al. (1981) found in their experiment on the readability of self-care instruction pamphlets, that the scores obtained for a text may differ by as much as 41.2% from the average score obtained for that text. Mailloux et al. (1995) compare programs for the calculation of FRE, FGS, and the Fog index and find that the three formula obtain significantly different scores on the same texts, FGS providing the lowest and Fog index the highest grade scores (Mailloux et al. 1995, p. 222).

While these issues do not disqualify readability formulae from being used, it should have become clear, that they are to apply and interpret with care: They are not necessarily equally applicable to all reading settings. Furthermore, the score scaling often suggest a straight forwards interpretation, yet, for their scale-external interpretation this might not be true. However, keeping these aspects in mind and choosing readability formulae accordingly, they provide an efficient way to predict readability of texts. Please see also Bailin and Grafstein (2001, p. 292) for further discussion.

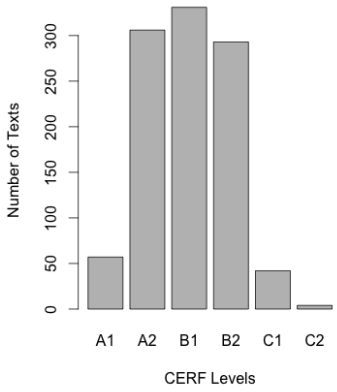


Figure 1: Histogram of grade levels in the German section of the Merlin corpus (N = 1,033).

4. Data Set

For the proficiency classification experiment and the evaluation of the RC the German subsection of the trilingual Merlin corpus was used (Abel et al. 2013).[2] The full Merlin corpus consists of 2,286 texts produced by learners of German, Italian and Czech in the course of standardized language certifications. They were scored based on the Common European Reference Frame (CERF) level scale A1 to C2. The German section has a size of 1,033 texts (Goullier 2007; Abel et al. 2013). Most texts are formal or informal letters or emails. Texts designed for C1 learners consist of essays, articles, and reports. Aside from information of author age, gender and L1 information, each text comes with separate proficiency ratings for grammatical accuracy, orthography, vocabulary range, vocabulary control, coherence and cohesion and sociolinguistics, as well as an overall score. However, the corpus is not balanced for grade levels, as may be seen in figure 1. This issue is addressed in subsection 6.1 The Merlin corpus has been used for the task of proficiency assessment before, employing elaborate sets of features of linguistic complexity, see for example Hancke, Vajjala, and Meurers (2012), Hancke (2013a, 2013b).

5. Program

The Readability Calculator (RC) was designed in order to obtain a variety of readability formulae. The code was written in Python version 3.4.3 and implements overall twelve readability formulae, which are discussed in subsection 5.2. The program is freely available for usage at GitHub.[3] After download, it may be run via command line with the command

```
> python3 main.py <input> <output> (counts)
```

2 http://merlin-platform.eu.
3 https://github.com/zweiss/RC_Readability_Calculator.

Where

input has to be substituted with the full path to the input directory containing the plain text files to be processed. All files ending in *.txt* in this directory and all its subdirectories will be processed by the program. Any meta information should be encoded in the file name.

output has to be replaced with the name the output file shall have, for example *output.csv*. In this file, the program writes the document $\times$ formulae/feature/count matrix in form of a comma separated value (CSV) file.

counts is an optional parameter. If it is omitted, the RC simply calculates all readability formulae. If it is given to the program, additionally to the regular output, it also produces lists of what was counted as a specific feature component. Subsection 5.4 elaborates on this verification option.

In order to run the program locally, an installation of Python 3 has to be installed, as well as the NLTK package.

5.1 Motiavation

It could be argued that there is already a multitude of readability formula calculation platforms available. So why programming another one, especially one distributed in form of plain Python code instead of coming with an easy to use web interface? While it is certainly true, that there already are several platforms offering the automatic analysis of texts with various readability formulae with or without charging fees, there is a certain shortage of platforms giving insights in the means by which they calculate the formulae and in reports on their robustness. This is an issue, because the same text given to different platforms for readability analysis returns very different results (cf. DuBay (2004)).

To illustrate this issue, a random text from the Merlin corpus was given to four different platforms offering a free readability analysis, two of which were designed for German and three for English texts. The platforms are readability-score.com[4], readabilityformulas.com[5], and online-utility.org[6] for English and textinspektor.de[7] as well as leichtlesbar.ch[8] for German.[9] The obtained sentence numbers range from 14 to 26 sentences,[10,11] the number of words counted from 155 to 162.[12] Also, while a cross-language comparison of syllable counts is not reasonable due to anticipated language differences in the constructions of syllables, it is interesting to compare the number of syllables observed on both German platforms, which is 308 for textinspektor.de and 270 for leichtlesbar.ch. Most remarkable might be a platform internal inconsistency observable at readabilityformulas.com, though, where an average number of syllables

4 `https://readability-score.com/`.
5 `http://www.readabilityformulas.com/`.
6 `http://www.online-utility.org`.
7 `http://www.textinspektor.de/`.
8 `http://www.leichtlesbar.ch/`.
9 Please note that this was neither intended nor designed as a proper experiment. Instead, it is meant as a first orientation concerning the available platforms.
10 By lichtlesbar.de and readability-score.com, respectively.
11 The extreme range in sentence number may partially be explained by the text type: As most texts in the Merlin corpus, the example was written in form of a letter and the address field imposes a challenge to sentence segmentation tools, see also section 5.4. Depending on the segmentation strategy, this results in very different sentence counts.
12 readability-score.com vs. textinspektor.de, readability-score.com, and readabilityformulas.com.

per word of 2 is reported, which seems dubious given their total counts of 260 syllables and 162 words. As $\frac{260}{162} = 1.6 \neq 2$, there seems to be some error in the feature calculation process, independent of the – unknown – robustness of the feature counts themselves.

Given these considerable differences, it is not surprising that the obtained scores differ across platforms even for the same formula, such as the Fog score, which ranges from 5.7 to 9.8 for the English platforms. As neither of the platforms report the robustness of their feature calculation nor elaborate on the NLP tools used to obtain their counts, it is not possible to reason about these differences and to decide which score to trust. Also, it is not possible to investigate the erroneous scores for average word length in syllables produced by readabilityformulas.com, since the formula implementation is not documented.[13] This lack of consistency and information is addressed by this paper and the sharing of the entire code.

5.2 Formula Selection

Overall, the RC features twelve different readability formulae, some of which are parametrized for German, some of which are parametrized for English. They were selected mainly based on their popularity, and how suited they seemed for the task, i.e. there was a preference for formulae known to be used for German and for formulae parametrized for L2 contexts. However, free formula availability was also a major factor out of necessity.[14] For conceptual purposes, it was decided to only use very simple, common readability formulae such as those mentioned in Section 2. Furthermore, formulae relying on vocabulary lists were excluded for reasons of experiment design in order to allow for an easier comparability of the formulae in the experiments: If all features rely exclusively on length measures of some sort, the effect of concrete length measure selection and parametrization may be investigated more clearly.

5.2.1 Flesch Scores. Three of the readability formulae implemented are based on the – especially in journalism highly influential – work of Rudolf Flesch (DuBay 2004, p. 22): the original Flesch Reading Ease (FRE) and two re-parametrized versions, the Flesch Grade Scale (FGS) and the Amstad Lesbarkeits Index (ALI). They all share the same general formula, as illustrated in equation 1.

$$FS = C - \alpha * SLW - \beta * WLS - \epsilon, \tag{1}$$

13 One theory, which might explain their puzzling result would be a code internal type-token mix-up, as readabilityformulas.com also reports the number of unique words, which was 122 for the given text and $\frac{160}{122} = 2.1$. At least, this is closer to the reported result. However, it would still imply a rounding error. It is left to the inclined reader to take up further investigations on this matter.

14 For example, the Hohenheimer Verständlichkeitsindex would have been highly interesting for the purposes of the experiments in terms of popularity and target language, yet the formula is not publicly available, which made a re-implementation in the RC impossible. See `https://bolzhauser.de/unsere-verfahren/hohenheimer-verstaendlichkeitsindex/` for more details.

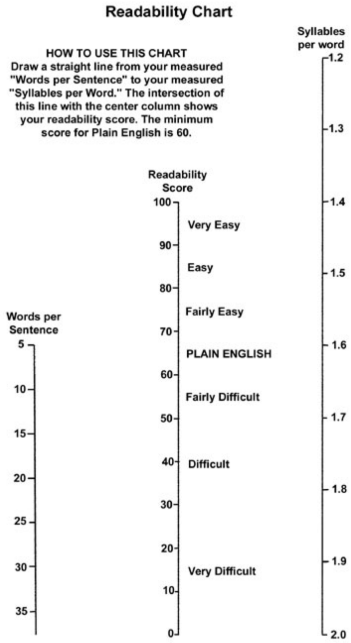


Figure 2: FRE interpretation scale from Flesch (2016).

with SLW being the average **s**entence **l**ength in **w**ords and WLS being the average **w**ord **l**ength in **s**yllables.[15] C, α, β, and ϵ are parameters and differ depending on the Flesch formula variant.[16]

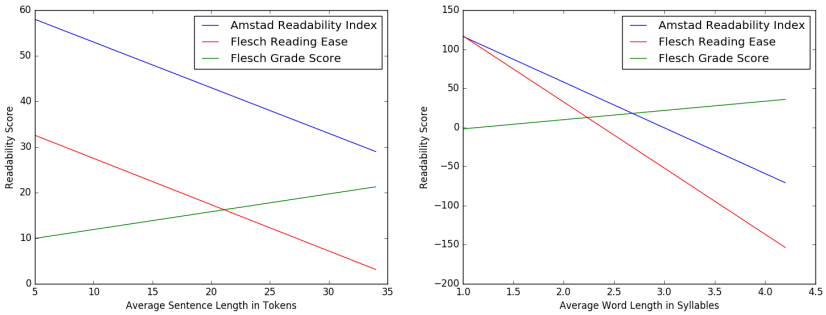
The original **Flesch Reading Ease** by Flesch (1948) is one of the most prominent and allegedly most reliable readability formulae (cf. DuBay (2004), Chall (1958), Klare (1963, p. 20f)). Its parameters are $C = 206.835$, $\alpha = 1.015$, $\beta = 84.6$, and $\epsilon = 0$. The original reading setting for these parameters is adult English L1 readers of news paper articles. Strictly speaking, the observable scores range between $[-\infty, 205.84]$, assuming that $ASW \geq 1$ and $WLS \geq 0$, i.e. imposing the minimal requirement that in order to be score-able texts have to consist of at least one single-token sentence, which does not necessarily have to contain a syllable.[17, 18] However, for most practical purposes it ranges between $[0, 100]$. figure 2 shows how to evaluate the resulting scores for this range. More extreme values are assumed to belong to the peripheral categories.

15 Please note that, in order to stay consistent with the literature on readability formulae, the linguistically unclear notion of "word" was kept throughout this paper. Unless explicitly defined otherwise, this corresponds to the regular definition of tokens.

16 It should be noted, that strictly speaking, neither of the formulae has been originally formulated as shown in equation 1. Yet, all of them may be re-formulated using the parameters given above, in order to match this generalized formula.

17 Possible zero-syllable single-token sentences are, for example, single utterances of numbers, i.e. *0405*, *0712* or sentences consisting of single punctuation marks, such as *!*, which may or may not occur depending on the sentence definition and data used. Plausible contexts for these examples are for example comics or chat language.

18 This minimal requirement assumption will be used to calculate score ranges throughout the paper.



(a) SLW = X, WLS = 2. (b) SLW = 5, WLS = X.

Figure 3: Flesch scores for varying features values.

The **Amstad Lesbarkeits Index** is a re-parametrization of the FRE for German, i.e. its reading setting is adult German L1 readers. It was developed by Amstad (1978) and uses the parameters $C = 180$, $\alpha = 1$, $\beta = 58.5$, and $\epsilon = 0$. The re-parametrization keeps the resulting scores within the anticipated range of figure 2, adjusting for the on average higher sentence- and word-lengths obtained in (Swiss-)German news paper articles. So while the scores may possibly range between $[-\infty, 179.00]$ (again assuming $SLW \geq 1$, $WLS \geq 0$), in practise they also tend to range between $[0, 100]$.

In 1975 the **Flesch Grade Scale**, also known as the Flesch Kincaid Grade Level, was developed. It is a re-parametrization of the FRE intended to map the obtained scores directly to grade levels, i.e. a score of 6 indicates that a text may be read with the reading abilities acquired in sixth grade (Kincaid et al. 1975). The parameters are $C = 0$, $\alpha = -0.39$, $\beta = -11.8$, and $\epsilon = 15.59$. The scores range between $[-15.2, \infty]$. By changing the polarity of the parameters α and β, the polarity of the score is inverted, too, i.e. for the FGS higher scores indicate decreasing readability. Accordingly, the reading setting are school texts read by children, with English as their L1.

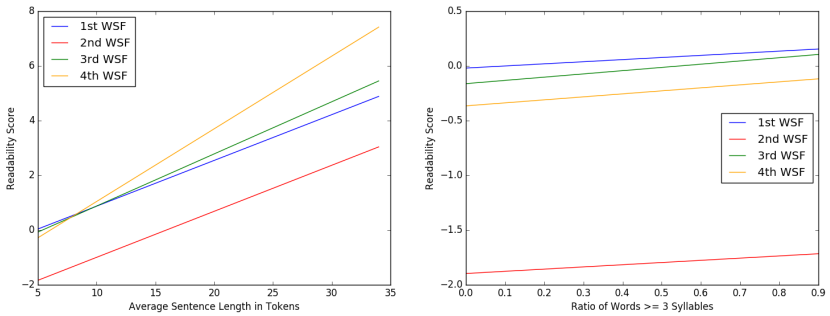
The differences between the scores returned by these three types of Flesch formulae are illustrated figure 3. Sub figure 3a shows the development of the three scores for varying sentence lengths, given an average word length fixed to 2. In sub figure 3b the development of the scores given a variable word length and a sentence length fixed to 5 is given. Aside from the difference in polarity between FGS and FRE visible in both figures, sub figure 3a clearly shows how the ALI is less sensitive to increasing word length than the FRE. However, in sub figure 3b it can be seen easily how ALI and FRE are perfectly correlated in terms of their reaction to increasing sentence length, except that the ALI returns scores nearly twice as high, i.e. it interprets longer sentences as less difficult than the original formula does. These two differences are as to be expected, given the re-parametrization for German.

5.2.2 Wiener Sachtext Formel. The **Wiener Sachtext Formel** was developed by Bamberger and Vanecek (1984) specifically for German factual texts read by adult L1 speakers. Four variants of the formula have been designed, using different parametrizations of the generalized formula in equation 2.

$$WSF = \alpha * LWS + \beta * SLW + \gamma * LWC - \omega * SWS - \epsilon, \quad (2)$$

WSF	α	β	ω	γ	ϵ
1st	0.1935	0.1672	0.1297	0.0327	0.875
2nd	0.2007	0.1682	0.1373	0	2.779
3rd	0.2963	0.1905	0	0	1.1144
4th	0.2744	0.2656	0	0	1.693

Table 1: Parameters for the four variants of the Wiener Sachtextformel (WSF).



(a) SLW = X, LWC = 0.3. (b) SLW = 5, LWC = X.

Figure 4: WSF scores for varying feature values.

where *LWS* is the average number of **l**ong **w**ords with three or more **s**yllables, *LWC* the average number of **l**ong **w**ords with six or more **c**haracters, and *SWS* the average number of **s**hort **w**ords with only one **s**yllable, and *SLW* as defined for the Flesch scores. table 1 shows the different parameter values for each of the four variants of the WSF. As may be seen, only the first WSF uses the average number of words with six or more characters as a feature, and the third and fourth WSF ignore single-syllable words. As for the FGS, the readability score returned is scaled such that it matches school grade levels. Accordingly, for most purposes, it ranges between $[4, 15]$, however, the actual limits ranges from $[-2.61, \infty]$ for the second WSF. The other formulae, too, have an upper limit of ∞, while their lower limit ranges between $[-1.43, -0.74]$.

Figure 4 shows the differences between the four formulae, once for varying sentence lengths with the ratio of long words fixed to 0.3, and once for varying ratios of long words with sentence length being fixed to 5. As can be seen, the second WSF returns lower values than the others. Also, the third WSF is more sensitive to increases in sentence and word length than the first WSF, as its slope is steeper. The fourth WSF mainly reacts on sentence length.

5.2.3 Fog Index. The **Fog index** was published by Gunning (1968) and is highly popular due to its simplicity and the high correlations with tests for readability formula verification (cf. DuBay (2004, p. 24)). It is given in equation 3.

$$GFI = 0.4 * (SLW + PSW), \tag{3}$$

with *SLW* as before and *PSW* as the number of polysyllabic words with three or more syllables. This parametrization was designed for a reading setting of adult English L1 readers, mainly of news paper articles, but also for other reading material (Gunning 1969). The scores range between $[0.4; \infty]$ and are negatively correlated with readability.

5.2.4 Coleman Liau Index. The **Coleman Liau Index** by Coleman and Liau (1975) was designed to allow for obtaining a readability score without syllable information. Hence, unlike the previous formulae, the formula in equation 4 works without syllable counts and relies instead solely on character, word, and sentence counts.

$$CLI = 0.0588 * WLC * 100 - 0.296 * \frac{num\ sentences}{num\ words} * 100 - 15.8, \tag{4}$$

with *WLC* being the average word length in characters. The original motivation for excluding syllable was the difficulty of accessing syllable information on texts digitalized with optical scanning devices Coleman and Liau (1975, p. 283). Clearly, this argument does not hold any more for the current state of the art, yet, in the interest of offering a broad variety of readability formulae the CLI was added to the RC nonetheless. Also, Smith and Senter (1967) argued for the validity of this approach in the context of the Automated Readability Index (ARI) (see below) due to the high correlation they found between syllable and character numbers in words (Smith and Senter 1967). The CLI ranges from $[-39.52, \infty]$ and is parametrized for various text types targeted at adult readers with English as their L1.

5.2.5 Lix Readability Index. The **Lix readability index** (Swedish: ***l**äsbrhets**i**nde**x***) was developed by Björnsson (1983) and is displayed in equation 5.

$$Lix = SLW + LWC, \tag{5}$$

again with *SLW* as the average sentence length in words and *LWC* as the average number of words consisting of six or more characters. While the two features themselves are quite common among readability features, the absence of parameters is rather uncommon for readability formulae. Yet, this is what is found to fit best for a reading setting of adult L1 readers for various languages by Björnsson (1983), who tested the Lix on Swedish, Danish, Norwegian, English, French, German, Italian, Spanish, Portuguese, Finnish, and Russian. As may be seen from the formula immediately, scores range from $[1; \infty]$ with higher scores indicating poorer readability.

A table with the concrete readability standards suggested by Björnsson (1983) for Swedish is shown in table 2. For German he reports that the Lix is on average 5 units higher than for Swedish, while Esfahani et al. (2014), Luers et al. (2013) report a Lix score of 46 as being well readable for German online texts. Yet, they do not provide a clear reference for this assessment. For potential adjustments of the Lix standard table for other languages, please see Björnsson (1983).

5.2.6 Automated Readability Index. The **Automated Readability Index** was designed by Smith and Senter (1967). As the CLI, it measures average word length in terms of characters, not in syllables, as is illustrated in equation 6.

$$ARI = 4.71 * WLC + 0.5 * SLW - 21.43, \tag{6}$$

Very easy text	20		
		25	Books for children
Easy text	30		
		35	Fiction
Average text	40		
		45	Factual prose
Difficult text	50		
		55	Technical literature
Very difficult text	60		

Table 2: Lix standards for Swedish (table 4 Björnsson 1983, p. 484).

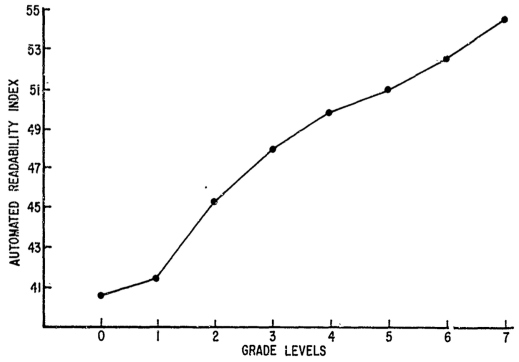


Figure 5: Relationship between grade level and ARI score, from figure 5 Smith 1967, p. 9).

with *WLC* and *SLW* as in the previous formulae. The scores range between $[-16.22; \infty]$

Smith and Senter (1967) obtained the weights for the ARI via regression analysis, where they tried to predict the grade levels assigned to text from schoolbooks with their formula (Smith and Senter 1967, p. 7f), i.e. the formula assumes a reading setting of children with English as L1. The levels ranged from grade 1 to grade 7. However, although Smith and Senter (1967, p. 9) correlate the ARI scores with grade levels, they point out the possibility of high variance across scores for different school systems. figure 5 shows the relation between the computed ARI and the grade levels presented by Smith and Senter (1967, p. 9).

5.2.7 Miyazaki EFL Readability Index. The Miyazaki EFL Readability Index (ML2RI) was developed by Greenfield (1999) in order to measure the L2 English proficiency of Japanese students. It was parametrized for academic texts for adult EFL Japanese readers. However, Greenfield (2003, p. 44) argues, that it should also apply to other L1 backgrounds. The ML2RI is the only L2 readability formulae, that is not only freely accessible, but also easy enough to fit to the other readability formulae. In fact, most more recent readability formulae as presented by Crossley, Greenfield, and McNamara (2008), François and Fairon (2012) actually qualify as proper models on linguistic complexity. Their design goes beyond fitting simple regression equations. Hence, they were disqualified for this collection of plain readability formlae. Equation 7 shows the ML2RI.

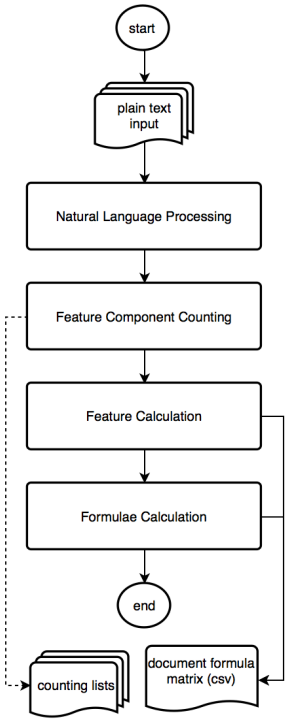


Figure 6: The RC workflow.

$$ML2RI = 164.935 - 18.792 * WLC - 1.916 * SLW, \tag{7}$$

with WLC and SLW as defined before. This score is again negatively correlated with readability, i.e. a decreasing score implies increasing readability. It ranges from $[-\infty; 144.23]$. According to Greenfield (2003, p. 43) a score of 50 indicates “average difficulty".

5.3 Pipeline

The RC’s pipeline is illustrated in figure 6. When executing the program, each input file is processed as follows: First, NLP tools are used in order to obtain token and sentence boundaries. For this task, implementations of NLTK version 3.2.1’s tokenizers are used: sentence segmentation is performed with the tokenizer from NLTK’s data module using the module’s German punctuation model.[19] For tokenization, the WordPunctTokenizer from NLTK’s tokenize module is used. After the NLP step, the following counts are obtained: number of sentences, number of tokens, number of tokens that are not punc-

19 See “tokenizers/punkt/german.pickle".

tuation marks,[20] number of characters, number of syllables,[21] number of words with one syllable, number of words with one or two syllables, number of words with three or more syllables, and number of words with six or more characters.The obtained counts are used to compute the following features in the next module: average sentence length in words, average word length in syllables, average word length in characters, average number of words with one syllable, average number of words with one or two syllables, average number of words with three or more syllables, average number of words with six or more characters, and ratio of sentences to words. With these features and counts, the twelve readability formulae reported in the previous section are calculated. Counts, features and readability formulae are saved in a document $\times$ count /feature/formula matrix in a CSV file.

5.4 Count Robustness

One important issue with readability formulae is that scores are known to differ depending on which implementation of a formula is used. To get varying scores for a given text using the same readability formula depending on the tool that was implementing it, is clearly not desirable. This is especially true, if the source of the difference cannot be investigated easily by inspecting the feature components on which the formulae are based, as discussed in subsections 5.1.

Because of this, the RC allows for an optional output of the feature component instances that were counted for each document. If the RC is executed using the optional *count* parameter, it creates one directory for each feature component in the given input directory. In these directories, it creates for each document an indexed list of all instances that have been counted for this feature component. Afterwards, the user may inspect these lists in order to inspect the robustness of the RC for a given data set. This semi-automatic approach may either be used for an informal inspection or to annotate a gold standard corpus to evaluate performance based on the number of true positives, false negatives and false positives for a given data set.

In order to quantify the robustness of the RC on the Merlin corpus, a random sample of one text per proficiency level ($N = 6$) was drawn from the Merlin data and then semi-automatically annotated with the approach described above. table 3 shows the amount of true positives, false positives, and false negatives found for the semi-automatically evaluated Merlin subset as well as precision, recall and f1-score. As may be seen, the RC is performs overall very well on the data set. Observed issues are especially due to the two NLTK tokenizers: in sentence segmentation, dots in abbreviations and dates are not always identified correctly. This is especially an issue on the Merlin corpus, since the majority text type are letters: Address fields and letter headers challenge sentence segmentizer, especially with respect to the issues with the processing of dates. Also, tokenization handles dates and abbreviations incorrectly, too.

20 Punctuation marks are considered to be periods, colons, commas, semi colons, hyphens, exclamation and question marks, double as well as single quotation marks, back and forward slashes, and round, angle, squared and curly brackets.

21 Syllables are obtained by counting the number of vowels, counting diphtongs only once. Hiatus, i.e. adjacent vowels in a word, that are not within the same syllable, such as in the German word *reell*, can not be identified, see Weiß (2015, p. 30) for further information on this issue.

Count	TP	FP	FN	Precision	Recall	f1-score
tokens	822	8	0	0.990	1	0.995
sentences	61	3	3	0.953	0.953	0.953
words $\geq$ 3 syllables	121	0	1	1	0.992	0.996
words $\leq$ 2 syllables	555	1	0	0.998	1	0.999
words with 1 syllable	324	0	0	1	1	1
words $\geq$ 6 characters	255	0	0	1	1	1
non punctuation tokens	709	0	0	1	1	1

Table 3: Performance of the RC on the Merlin sub corpus ($N = 6$).

6. Experiment

This section reports on the classification experiments performed to predict the overall CERF level attributed to L2 German speakers with readability formulae. It should be noted, that this paper purposely avoids mapping readability scales to CERF levels, though: As has been discussed previously, the use of presumably easy to interpret readability scales fosters misleading interpretations and should, hence, not be encouraged. The code used to perform all experiments may be found in *ml_experiment.py* at `https://github.com/zweiss/RC_Readability_Calculator/rc_code/`.

6.1 Method

For each readability formula a separate classification experiment was performed trying to predict the overall CERF score. Additionally, one experiment was performed using all features calculated for all readability formulae as predictors to establish a higher-order baseline. Each of the experiments was performed with multinomial logistic regression. For this the *linear_model* module of the *sklearn* package was used. Furthermore, a majority baseline was established for comparison using the *DummyClassifier* from the *dummy* module in the *sklearn* package. Overall, this leads to a total amount of fourteen distinct classification experiments.

In order to evaluate the performance and address the issue of class imbalance, stratified 10-folds cross validation was performed using *sklearn*'s *cross_validation* module. Stratification in this contexts means, that the algorithm tries to design balanced folds. This does only allow for *n*-fold cross validation, where $n \geq$ the minimal amount of class instances in the response variable. Hence, the C2 CERF level had to be dropped: The Merlin corpus features four instances of German C2 writings, only, which is not enough to allow for a sufficient amount of balanced folds. However, it seemed more desirable to drop these four instances than to either decrease the amount of folds or to waive stratification. Significance of the differences between f1-score, precision, and recall is evaluated across formulae in comparison to three reference levels: On the one hand, there are the basic majority baseline and the more elaborate feature baseline. Hence, formulae may be evaluated in terms of whether they a) perform better than chance and b) their feature selection and weights are more informative than all features together. On the other hand, they are also compared with the overall best performing formula (CLI), in order to test whether CLI performs significantly better than the other formulae. For significance testing, a two sided t-test was employed, since all predictors are based on the same text population, and a conservative significance level of $\alpha \leq 0.01$ was chosen.

Formula	f1-score	Precision	Recall
Majority baseline	**0.157 –, v, v**	**0.104 –, v, v**	**0.322 –, v, v**
WSF 4	0.420 *, ^{v}, v	0.417 *, ^{v}, v	0.449 *, v, v
Fog	0.420 *, ^{v}, v	0.418 *, ^{v}, v	0.450 *, v, v
WSF 3	0.423 *, ^{v}, v	0.421 *, ^{v}, v	0.453 *, v, v
Lix	0.425 *, ^{v}, v	0.422 *, ^{v}, v	0.454 *, v, v
WSF 2	0.427 *, ^{v}, v	0.424 *, ^{v}, v	0.457 *, v, v
WSF 1	0.428 *, ^{v}, v	0.425 *, ^{v}, v	0.458 *, v, v
Feature baseline	**0.484 *, –, v**	**0.484 *, –, v**	**0.522 *, –, v**
FGS	0.488 *	0.488 *	0.527 *
ML2RI	0.488 *	0.487 *	0.528 *
ARI	0.488 *	0.488 *	0.528 *
ALI	0.489 *	0.485 *	0.529 *
FRE	0.494 *	0.491 *	0.533 *
CLI	**0.515 *, *, –**	**0.514 *, *, –**	**0.554 *, *, –**

Table 4: Merlin CERF level prediction performance across readability formulae. It includes significance markers for majority baseline (1. position), formula baseline (2. position), and CLI (3. position). * indicates significantly higher, v significantly lower than comparison level for $\alpha \leq 0.01$. Superscript indicates $\alpha \leq 0.05$.

However, results obtained for a less strict significance level of $\alpha \leq 0.05$ are also indicated in table 4 and partially referred to later on for sake of completeness.

6.2 Results

The results of the classification experiments may be seen in table 4. Experiments are displayed in rows and ordered according to increasing f1-scores.

The majority baseline has a weighted average f1-score of 15.7% resulting from a weighted average precision of 10.4% and an about three times higher weighted average recall of 32.2%. For all three performance measures all readability formulae as well as the feature baseline return significantly better results. Also, while precision continues to be lower than recall throughout all experiments, the extend of the difference decreases noticeable.

The feature baseline shows a weighted precision average f1-score of 48.4%, while having a 3.8% higher recall. Unlike the basic majority baseline, the feature baseline imposes a higher threshold to the readability formulae, which is not exceeded in every experiment: Except for CLI, none of the formulae performs significantly better on any performance measure than the feature baseline. In fact, all WSFs as well as the Lix and Fog index have a significantly lower recall. Assuming a less strict significance level of $\alpha \leq 0.05$ this extends to f1-score and precision as well.

The CLI performs overall best in terms of all three readability measures: It shows significantly higher precision, recall and f1-scores compared to both baselines, all WSFs, the Fog and the Lix index. It displays one of the smallest differences between precision and recall. However, there is no significant difference to either of the Flesch based formulae, the L2 readability formula or the ARI.

As for the performance of related formulae, it should be noted that neither of the Flesch formulae significantly outperforms the other. The same holds for the WSFs.

6.3 Discussion

Throughout all classification experiments, recall was shown to be higher than precision. While this holds especially for the majority baseline, where a difference of 21.8% is observable, also for all other formulae recall is at least about 3% higher than precision. Given the class unbalance in the Merlin corpus, this is not surprising. It merely shows how classification consistently favours the three most frequent classes A2, B1, and B2, while not classifying A1 and C1 properly. However, it is remarkable how much the readability formulae as well as the feature baseline improve on this issue. This indicates that while class imbalance continues to challenge classification, the formulae manage to also include minority classes.

This as well as the significant increase in precision, recall, and f1-score to be seen in all experiments, shows that readability formulae do perform above chance on the task of proficiency assessment based on written language productions. This holds even for those applied to a mainly unfitting 'reading' setting, such as the ML2RI, which is designed for Japanese EFL readers or the FGS, which is designed for American students in the 1980s reading their native language.

This becomes even more clear, when considering the feature baseline. Four of five readability formulae designed specifically for German perform significantly worse than the feature baseline for $\alpha \leq 0.05$. This is especially unexpected, as they are all variants of the WSF, which was not only designed for German but also uses a comparatively large amount of linguistic information. Fog and Lix index, however, which perform equally bad, are also the two most simple formulae implemented. Thus, their performance might be due to the lack of feature weighting, while the performance of the WSFs may not. Two possible explanations occur: First, the weights set for the formulae might be not suited for the task of proficiency assessment. Second, the formulae might employ too many features, since – as discussed earlier – it is often noted that more features used in formulae might in fact harm their performance. However, the first assumption seems to be more plausible: Otherwise WSF 3 and 4 should outperform the other two formulae, as they employ less features, but this is not the case.

Another interesting point is the general performance of the feature baseline: It employs overall eight raw features without any previously defined weights and either outperforms most other formulae or is not significantly different from them. This might be explained assuming that more information is in fact beneficial, if it is not paired with previously assigned weights.

In the light of this reasoning and the other results, the performance obtained for the CLI is truly remarkable: It is the only readability formula, which is significantly outperforming both baselines as well as half of the readability formulae. Yet, it is also one of the formulae using the least linguistic information, as it does not rely on any syllable information. Instead, it works based on the average number of characters per word and the sentence word ratio, only, combining those with a comparatively high amount of weights. Furthermore, the CLI was designed for English L1 academic texts, i.e. it operates on a completely different reading setting. Unfortunately, it goes beyond the scope of this paper to investigate this issue further.

6.4 Limitations

Readability formulae are usually not used by readers to estimate whether to read a text or not, but by writers. That is, they function as writing aides. As such, when performing a transfer to the task of proficiency assessment, it seems to be in order to comment on the actual applicability of readability formulae in this area, especially in the light of the obtained results.

For this, it should first be noted, that it is debated whether readability formulae are suited as L1 writing aides. Critics argue, they lead writes to blindly reduce sentence and word length, which does not necessarily lead to an increase in text quality (cf. DuBay (2004, p. 37)). Flesch (2016) argues contra to that, that proficient writers in fact automatically address various aspects of linguistic complexity when being told to reduce sentence length. No matter which stance in this debate one takes, it is obvious that this does not necessarily hold for L2 writers: It is not to be expected, that by simply writing longer sentences and words, text quality will improve. Depending on the proficiency level, in fact, the opposite might be the case: syntactical and morphological errors might increase.

In order to provide useful aide for L2 writers, linguistically more detailed feedback and information has to be provided. However, as has been established, readability formulae are insensitive to linguistic details. They only superficially correlate with complex linguistic construction, which is what makes them so easy to obtain in the first place. Clearly, this form of feedback is only useful in a very limited sense, such that it allows for a quick and easy, but rough estimation of the current proficiency level. It should not be used to actually grade learners and might only be useful for self-learning reading in a very restricted sense and in combination with other learning aides.

7. Future Work

While the RC already implements a broad range of readability formulae, it should be pointed out, that some restrictions in terms of variety had to be accepted. On the one hand, only simple readability formula using length measures have been considered. This allowed for a clearer interpretation of the obtained results, yet, a comparison to formulae using word lists might also be interesting. On the other hand, only a single readability formula tested for L2 proficiency assessment was in fact designed for L2 contexts. Their lack of popularity and their rare number makes it difficult to obtain L2 formulae: most contemporary work on the topic developed in the direction of more complex models of linguistic complexity, which are not reduced to simple regression equations and rely on a way larger set of features, see for example Crossley, Greenfield, and McNamara (2008), François and Fairon (2012). However, regarding the promising performance of the ML2RI it would certainly be desirable to enhance the RC with even more L2 formulae.

Aside from this, it has to be stated that not all interesting questions arising from the obtained data could be answered exhaustively due to the limited scope of this paper: Further work on the relation between learner L1 and formula performance in the Merlin corpus might be interesting. Also, the differences of scores assigned by readability formulae using similar scales could not be discussed in further detail: As this paper specifically focuses on the predictive power of readability formulae in proficiency contexts, and a mapping of for example grade scores to CERF levels does not seem desirable, the detailed investigation of the actual scores assigned did not fit in this context. This is not to say, that it would not be a question worthwhile to investigate

separately. The same holds for an analysis of how the results obtained by the RC relate to results obtained by other readability analysis platforms.

Last but not least, it would be desirable to compare the performance of readability formulae with the performance of more elaborate feature models of linguistic complexity, such as those presented in Hancke, Vajjala, and Meurers (2012), Hancke (2013a, 2013b). These models employ up to 200 features in order to determine text readability and writer proficiency. However, they are also heavily linked to various other tasks in the field of computational stylometry. In the contexts of readability and proficiency assessment, a comparison of formulae and elaborate state of the art models of linguistic complexity would be very interesting. While the previously mentioned issues and questions might be addressed at a later point, this comparison will be investigated in the near future.

8. Conclusion

This paper investigated the applicability of readability formulae to the domain of L2 proficiency assessment. The main question was, whether such a transfer is possible. F1-scores up to 51.5% for a five-partite classification problem were obtained. Thus, it may be stated, that transfer is in fact possible for some formulae. This is especially interesting, since the CLI, which performed best, is in fact neither designed for German nor for L2 contexts in the first place. These results show, how notoriously unclear the performance of readability formulae on varying reading settings is: it is neither clear, that a mainly fitting reading setting leads to good results, otherwise the WSFs should have returned better results, nor that mainly unfitting reading settings lead to bad performance, otherwise the CLI should not have returned the best results. So a transfer of domain seems to be possible in terms of predictive power, and interesting with regard to the influence of reading setting. However, for practical purposes readability formulae do not seem to be suited for L2 learning contexts, due to their limited exploratory power.

References

[Abel et al.2013]Abel, Andrea, Lionel Nicolas, Jirka Hana, Babora *Štindlov*, Serhiy Bykh, and Detmar Meurers. 2013. merlin: A trilingual learner corpus illustrating european reference levels. LRC 2013, Bergen, Norway, September.

[Amstad1978]Amstad, T. 1978. *Wie verständlich sind unsere Zeitungen?* Ph.D. thesis, University of Zürich.

[Bailin and Grafstein2001]Bailin, Alan and Ann Grafstein. 2001. The linguistic assumptions underlying readability formulae: a critique. *Language & Communication*, 21:285–301.

[Bamberger and Vanecek1984]Bamberger, Richard and Erich Vanecek. 1984. *Lesen – Verstehen – Lernen – Schreiben. Die Schwierigkeitsstufen von Texten deutscher Sprache*. Jugend und Volk, Wien.

[Björnsson1983]Björnsson, C. H. 1983. Readability of newspapers in 11 languages. *Reading Research Quarterly*, 18(4):480–497.

[Brettschneider2012]Brettschneider, Frank. 2012. Faqs im klartext-check: Statt antworten erhalten kunden oft nur kauderwelsch. Technical report, Press and publicity work of the University of Hohenheim.

[Brettschneider2013]Brettschneider, Frank. 2013. Kauderwelsch statt klartext: Kundenkommunikation von telekommunikations-unternehmen ist oft unverständlich. Technical report, Press and publicity work of the University of Hohenheim.

[Brettschneider2014]Brettschneider, Frank. 2014. Ceo-reden im vergleich: Ceo-reden im vergleich: Am verständlichsten ist bmw-chef norbert reithofer. Technical report, Press and publicity work of the University of Hohenheim.

[Brettschneider2016]Brettschneider, Frank. 2016. Langzeitstudie wahlprogramme: Parteien bemühen sich wieder zunehmend um verständlichkeit. Technical report, Press and publicity work of the University of Hohenheim.

[Brown1992]Brown, J. D. 1992. What text characteristics predict human performance on cloze test items? In *Proceedings of the 3rd conference on second language research in Japan*, pages 1–26.
[Brown1997]Brown, James Dean. 1997. An efl readability index. *University of Hawai'i Working Papers in English as a Second Language*, 15(2):85–119.
[Chall1958]Chall, J. S. 1958. *Readability: An appraisal of research and application*. Ohio State University Press.
[Chall and Dale1995]Chall, J. S. and E. Dale. 1995. *Readability revisited, the new Dale-Chall readability formula*. Brookline Books.
[Coleman and Liau1975]Coleman, Meri and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.
[Crossley, Greenfield, and McNamara2008]Crossley, Scott A., Jerry Greenfield, and Danielle S. McNamara. 2008. Assessing text readability using cognitively based indices. *Tesol Quarterly*, 42(3):475–493.
[DuBay2004]DuBay, William H. 2004. The principles of readability. *Online Submission*.
[Esfahani et al.2014]Esfahani, B Janghorban, A Faron, KS Roth, HE Schaller, F Medved, and JC Lüers. 2014. Systematic analysis of the readability of patient information on the websites of clinics for plastic surgery. *Handchirurgie, Mikrochirurgie, plastische Chirurgie: Organ der Deutschsprachigen Arbeitsgemeinschaft fur Handchirurgie: Organ der Deutschsprachigen Arbeitsgemeinschaft fur Mikrochirurgie der Peripheren Nerven und Gefasse*, 46(6):369–374.
[Flesch1948]Flesch, R. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233.
[Flesch2016]Flesch, Rudolf. 2016. How to write plain english. `http://www.mang.canterbury.ac.nz/writing_guide/writing/flesch.shtml`.
[François and Fairon2012]François, Thomas and Cédrik Fairon. 2012. An "ai readability" formula for french as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477, Jeju Island, Korea, July. Association for Computational Linguistics.
[Goullier2007]Goullier, Francis. 2007. The common european framework of reference for languages (cerf) and the developement of language policies: challenges and responsibilities. Technical report, Intergovernmental Language Policy Forum.
[Greenfield1999]Greenfield, Jerry. 1999. *Classic readability formulas in an EFL context: Are they valid for Japanese speakers?* Ph.D. thesis, Temple University.
[Greenfield2003]Greenfield, Jerry. 2003. The miyazaki efl readability index. *Comparative Culture*, 9:41–49.
[Greenfield2004]Greenfield, Jerry. 2004. Readability formulas for efl. *Japan Association for Language Teaching*, 26(1):5–24.
[Gunning1968]Gunning, Robert. 1968. *The technique of clear The technique of clear writing*. McGraw-Hill, New York.
[Gunning1969]Gunning, Robert. 1969. The fog index after twenty years. *Journal of Business Communication*, 6:3–13.
[Hamsik1984]Hamsik, M. J. 1984. *Reading, readability, and the ESL reader*. Ph.D. thesis, University of South Florida.
[Hancke2013a]Hancke, Julia. 2013a. Automatic prediction of cerf proficiency levels based on linguistic features of learner language. Master's thesis, Eberhard Karls Universität Tübingen, April.
[Hancke2013b]Hancke, Julia. 2013b. Exploring of cerf classification for german based on rich linguistic modeling. In *Learner Corpus Research Conference*, pages 54–56.
[Hancke, Vajjala, and Meurers2012]Hancke, Julia, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for german using lexical, syntactic and morphological features. In *Proceedings of COLING*, pages 1063–1080, Mumbai, December.
[Janghorban Esfahani et al.2015]Janghorban Esfahani, B., A. Faron, K. S. Roth, P. P. Grimminger, and J.-C. Luers. 2015. Systematische analyse der lesbarkeit von patienteninformationstexten auf internetseiten von kliniken für allgemein- und viszeralchirurgie deutscher universitätskliniken. *Online Submission*.
[Kennedy and McNally2005]Kennedy, Christopher and Lousie McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 81(2):345–381.
[Kincaid et al.1975]Kincaid, J. P., R. P. Fishburne, R. L. Rogers, and B. S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *CNTECHTRA Research Branch*, pages 8–75.
[Klare1963]Klare, G. R. 1963. *The measurement of readability*. Iowa State University Press.

[Leichter et al.1981]Leichter, Steven B, Janice A Nieman, Robert W Moore, Peggy Collins, and Anne Rhodes. 1981. Readability of self-care instructional pamphlets for diabetic patients. *Diabetes Care*, 4(6):627–630.

[Leong, Ewing, and Pitt2002]Leong, Elain K. F., Michael T. Ewing, and Leyland F. Pitt. 2002. E-comprehension. evaluating b2b websites using readability formulae. *Industrial Marketing Management*, 31:125–131.

[Lively and Pressey1923]Lively, Bertha A.. and Sydney Leavitt Pressey. 1923. *A method for measuring the "vocabulary burden" of textbooks*.

[Luers et al.2013]Luers, J.-C., A.-O. Gostian, K. S. Roth, and D. Beutner. 2013. Lesbarkeit von medizinischen texten im internetangebot deutscher hno-universitätskliniken. *HNO*, 61(8):648–654.

[Mailloux et al.1995]Mailloux, Steven L, Mark E Johnson, Dennis G Fisher, and Timothy J Pettibone. 1995. How reliable is computerized assessment of readability? *Computers in nursing*, 13:221–221.

[McLaughlin1969]McLaughlin, G. Harry. 1969. Smog grading – a new readability formula. *Journal of Reading*, May.

[Smith and Senter1967]Śmith, E. A. and R. J. Senter. 1967. Automated readability index. Technical report, DTIC Document.

[Vajjala and Meurers2012]Vajjala, Sowmya and Detmar Meurers. 2012. On improving the accuracy of read- ability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Innovative Use of NLP for Building Educa- tional Applications*, volume 7, pages 163–173, Montréal, Canada. Association for Computational Linguistics.

[von der Brück2007]von der Brück, Tim. 2007. A semantically oriented readability checker for german. In *Proceedings of the 3rd Language and Technology Conference*, pages 270–274, Poznań, Poland, October.

[von der Brück2008]von der Brück, Tim. 2008. A readability checker with supervised learning using deep indicators. *Informatica*, 32:429–435.

[Weiß2015]Weiß, Zarah Leonie. 2015. More linguistically motivated features of language complexity in readability classification of german textbooks: Implementation and evaluation. B.A. Thesis, September.