Applying Readability Formulae to the Task of Proficiency Assessment

Zarah Leonie Weiß

`zarah-leonie.weiss@student.uni-tuebingen.de`

Abstract

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% on the German subsection of the MERLIN corpus. It also introduces a freely available program for calculating readability formula for German texts, the Readability Formulae and Feature Calculator (RFFC), which is available at ???.

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 language (L1) and L2 learning contexts, on which a large amount of past and present research has focused (**greenfield???**; cf. Crossley, Greenfield & McNamara 2008; DuBay 2004). Yet, readability assessment is also highly interesting from an economical perspective for companies selling texts, such as news papers or books. They ha 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, see for example **flesch**; **gunning** on newspaper articles, but also Leong, Ewing & Pitt (2002) in the context of business to business communication. 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 (**esfahani**; see e.g. 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 with or without charging fees. Hence, their frequent use in many research contexts is hardly surprising.

Despite their popularity readability formulae are not uncontroversial, 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. ibid.). 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.

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, see for example **???**. 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. For this task, a new tool to obtain various readability scores for German based on plain text input was written using Python[1] and the Natural Language Tool Kit (NLTK)[2]: the RFFC is freely available at `???`. 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 4 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 excursion providing some background on readability formulae and their usage. Section 4 elaborates on the RFFC, 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. Then, the MERLIN corpus, which is used as data set for the classification experiment, is introduced. Afterwards, the classification experiments are reported and discussed in section 5. The paper closes with a note on limitations of the approach and a conclusion.

2   Related Work

The earliest readability formula was designed by **lively1923** (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 & Pitt (2002) use Dale-Chall grade level, Flesch Reading Ease (FRE), and Fog index to assess the readability of business-to-business web pages. **esfahani**; Esfahani et al. (2014) and

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 using five readability formula: Amstad Lesbarkeits Index (ALI), German SMOG index, the Wiener Sachtext Formel (WSF), and Hohenheimer Verständlichkeitsindex.

However, they 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 telecommunications 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.[3] For a comprehensive overview of 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, **brown1992** does

---

[3]The Hohenheimer Verständlichkeitsindex combines average sentence, clause, and word length with the ratio of words with more than six characters, of clauses with more than twelve tokens and of sentences with more than 20 tokens, cf. `https://bolzhauser.de/unsere-verfahren/hohenheimer-verstaendlichkeitsindex/`. However, the details of the formula are not publicly available, which made a re-implementation in the RFFC impossible.

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 **brown1992**'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 (ibid.) 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 (see Crossley, Greenfield & 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, mainly relying 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 Coleman Liau Index (CLI) by Coleman & 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, ???, Greenfield 2004, p. 7). Despite this conceptual argument, various studies have found them to correlate highly with other measures of readability. For example, Greenfield (ibid., 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 (ibid., 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, the criticism is correct insofar as 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 (ibid., p. 7), but also the correlation of syllable counts with morpheme counts and Yngve's word depth (DuBay 2004, p. 19).

It should be noted, though, 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 formulae may generalize well for certain aspects of this setting, it does not necessarily so for the other aspects. For example, Greenfield (2004, pp. 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 (ibid., p. 8) and DuBay (2004) for further considerations regarding changes in reading group. Unfortunately, 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 & Pitt 2002) or medical texts (**esfahani2013**; Esfahani et al. 2014; Luers et al. 2013), but also in the lack of disclaimers towards the readability setting of formulae on the various platforms offering readability analyses.

The issue of generalizability of readability formulae relates to the question of how to interpret obtained scores. 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 judgement adjectives (e.g. 'easy' and 'difficult'). While these allow for an easy scale-immanent comparison of texts, these 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 extremes stays notoriously difficult. For example, scores at the lower end of the FRE are argued to correspond to the reading level of American functional alliterates (**???**), allowing for a good interpretation of the 'very easy' index, while leaving the above stated questions concerning the 'fairly easy' index untouched. In general, the issue of vagueness of relative adjectives is well known and broadly discussed in the field of adjective semantics and further elaborations would unfortunately be beyond the scope of this paper. Please see **???** for further discussion.

Grade scaling avoids these problematic issues to a certain extend. 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 would return an average score of 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 largely. For example, scores calculated by the Dale-Chall formula by **???** are known to be generally two grades lower than scores predicted by the SMOG formula by **mclaughlin** (cf. DuBay 2004, p. 47). Also, 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. Finally, Mailloux et al. (1995) compare programs for the calculation of FRE, FGS, and the Gunning Fog index and find that the three formula obtain significantly different scores on the same texts, FGS providing the lowest and Gunning Fog index providing the highest grade scores (ibid., 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, their scale-external validity might not be given. However, keeping these aspects in mind and choosing readability formulae accordingly, they provide an efficient way to predict readability of texts.

4 Program

The RFFC was designed in order to obtain a variety of readability formulae. The code was written with Python version 3.4.3 and implements overall twelve readability formulae, which are discussed in subsection 4.2. The program is freely available for usage at GitHub.[4] After the download, it may be run via command line with the command

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

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 × formulae/count matrix in form of a comma separated value (CSV) file.

counts is an optional parameter. If it is omitted, the RFFC 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 for a specific feature component count. Subsection 4.4 elaborates on this verification option function.

In order to run the program locally, an installation of Python 3 has to be installed, as well as the NLTK package.[5]

4.1 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 `https://readability-score.com/`, `http://www.readabilityformulas.com/`, and `http://www.online-utility.org` for English and `http://www.textinspektor.de/` as well as `http://www.leichtlesbar.ch/` for German.

The obtained sentence numbers range from 14 to 26 sentences,[6] the number of words counted from 155 to 162.[7] Also, while a cross-language comparison of syllable counts is not reasonable due to 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 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.

Given these considerable differences, it is not surprising that the obtained scores differ across platforms even for the same formula, such as the a Gunning Fog score, which ranges from 5.7 to 9.8 for the English platforms. As neither of the platforms report their features robustness 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 reproduce the erroneous scores for average word length in syllables produced by readabilityfor-

---

[4]???.

[5]???.

[6]By lichtlesbar.de and readability-score.com, respectively.

[7]readability-score.com vs. textinspektor.de, readability-score.com, and readabilityformulas.com.

mulas.com, since the formula implementation is not documented.[8] This lack of information is addressed by this paper and the sharing of the entire Python code.

4.2 Formula Selection

Overall, the RFFC features twelve different L1 readability formulae, some of which are parametrized for German, some of which are parametrized for English. Three of them are based on Rudolf Flesch, four are variants of the Vienna formula for factual texts (*Wiener Sachtextformel*). The rest of the formulae is not related to each other. There is also one formula parametrized for L2 readers, however, it was designed for Japanese.

4.2.1 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.

$$FS = C - \alpha * SLW - \beta * WLS - \epsilon, \quad (1)$$

with SLW being the average **s**entence **l**ength in **words** and WLS being the average **w**ord **l**ength in **s**yllables.[9] C, α, β, and ϵ are parameters and differ depending on the Flesch formula variant.[10]

The original **Flesch Reading Ease** from Flesch (1948) is one of the most prominent and reliable readability formulae (cf. Chall 1958; DuBay 2004; Klare 1963, 20f). Its parameters are $C = 206.835$, $\alpha = 1.015$, $\beta = 84.6$, and $\epsilon = 0$, i.e. average word

Figure 1: FRE interpretation scale from **flesch2016**.

length is weighted stronger than average sentence length. The original reading setting for these parameters is adult English L1 readers. Strictly speaking, the observable scores range between $[-\infty, 205.84]$, assuming that $ASL \geq 1$ and $AWL \geq 0$, 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.[11] However, for most practical purposes it ranges between $[0, 100]$. Figure 1 shows how to evaluate the resulting scores for this range. More extreme values are assumed to belong to the peripheral categories.

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 $\epsilon = 0$, i.e. while word length is still the dominating factor, it receives less weight as well as sentence length. The

---

[8]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.

[9]Unless explicitly defined otherwise, the notion of "word" used in the readability formulae corresponds with the regular definition of tokens.

[10]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.
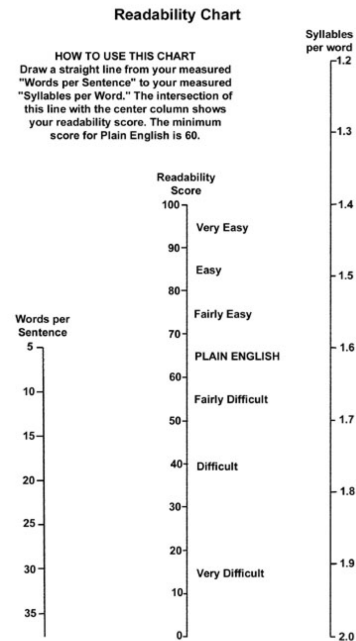
[11]Possible zero-syllable single-token sentences are, for example, single utterances of numbers, i.e. *0405* or sentences consisting of single punctuation marks, such as *!*, which may or may not occur depending on the sentence definition and data used. Plausible contexts for these examples are for example comics or chat language.
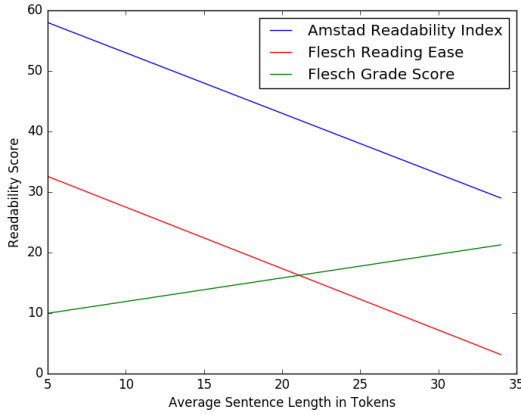
Figure 2: Flesch scores for varying sentence lengths assuming average word length of 2.

Figure 3: Flesch scores for varying word lengths assuming average sentence length of 5.

re-parametrization keeps the resulting scores within the anticipated range of Figure 1, adjusting for the on average higher sentence- and word-lengths obtained in German. 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 1978 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 (**flesch1978**). The parameters are $C = 0$, $\alpha = -0.39$, $\beta = -11.8$, and $\epsilon = 15.59$ and 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 is children, who are English L1 readers.

The differences between the scores returned by these three types of Flesch scores are illustrated in the follwing two figures: Figure 2 shows the development of the three scores for given varying sentence lengths and an average word length fixed to 2. In Figure 3 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, Figure 2 clearly shows how the ALI is less sensitive to increasing word length than the FRE. However, in
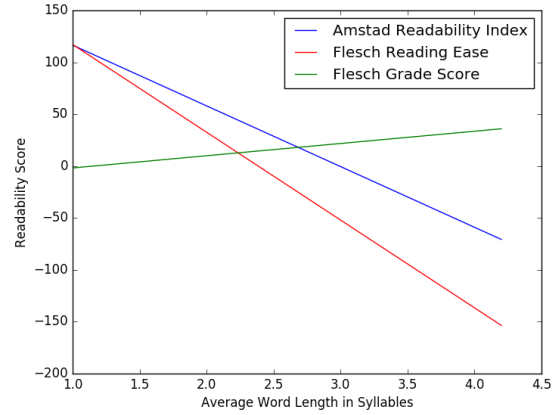
Figure 3 it can be seen easily how ALI and FRE are perfectly correlated in terms of their reaction to increasing sentence length. However, the ALI returns scores nearly twice as high, i.e. it interprets longer sentences as less difficult than the original formula does.

4.2.2 Wiener Sachtext Formel

The WSF was developed by **bamberger1984** specifically for German factual texts. There are four variants of the formula, using different parametrizations of the generalized formula in Equation 2.

$$
\begin{aligned}
WSF = \alpha * LWS + \beta * SLW \\
+ \gamma * LWC - \omega * SWS - \epsilon,
\end{aligned}
\tag{2}
$$

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 2 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 $[4, 15]$,

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

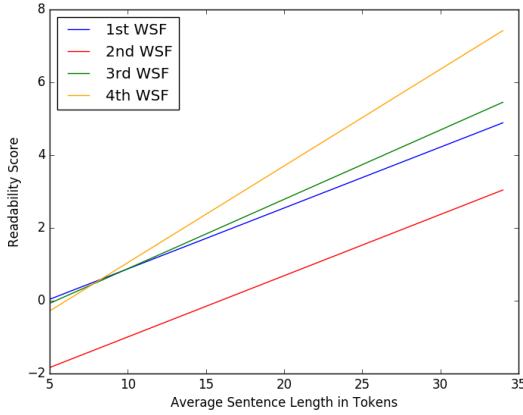Table 1: Parameters for the four variants of the Wiener Sachtextformel (WSF).

Figure 4: WSF scores for varying average sentence lengths assuming ratios of long words being fixed to 0.3.

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 $[-1.43, -0.74]$.

Figures 4 and 5 show 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. ******************************* The fourth formula shows TODO ******************************

4.2.3 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.

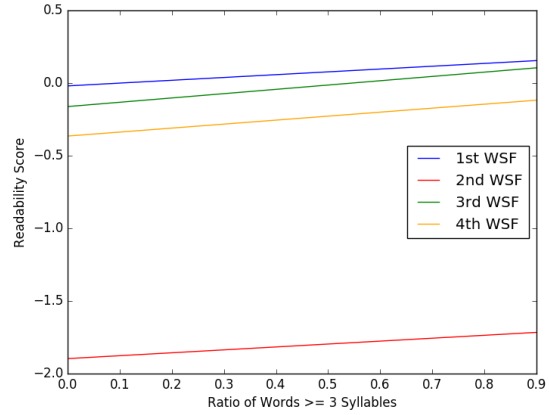$$GFI = 0.4 * (SLW + PSW), \qquad (3)$$

Figure 5: WSF scores for varying ratios of long words assuming average sentence length of 5.

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 ** reading setting: adult English L1 readers (educated guess) ** The scores range between $[0.4; \infty]$ and are negatively correlated with readability.

4.2.4 Coleman Liau Index

The CLI by Coleman & 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 and sentence counts.

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

with *WLC* being the average word length in characters and *sents*. The original motivation for excluding syllable was the difficulty of accessing syllable information on texts digitalized with optical scanning devices Coleman & Liau (ibid., 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 RFFC nonetheless. Also, Smith & 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 & Senter 1967).

The CLI ranges from $[-39.52, \infty]$ and is parametrized for *** reading setting: adult English L1 readers (educated guess) ***

4.2.5 Lix Readability Index

The Lix readability index (*läsbrhetsindex*, Swedish) 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, giving equal weight to both features, 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 (ibid.), 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 (ibid.) for Swedish is shown in Figure **??**. For German he reports that the Lix is on average 5 units higher than for Swedish, while **esfahani**; 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).

4.2.6 Automated Readability Index

The ARI was designed by Smith & Senter (1967). As the CLI, it measures average word length in terms of characters, not in syllables, as is illustrated

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 6: Relationship between grade level and ARI score, from **schmith1967**.

in Equation 6.

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

with *WLC* and *SLW* as in the previous formulae.

Smith & Senter (ibid.) 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 (**smith967**), 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 & Senter (ibid., p. 9) correlate the ARI scores with grade levels, they point out the possibility of high variance across scores for different school systems. Figure **??** shows the relation between the computed ARI and the grade levels presented by Smith & Senter (ibid., p. 9).

4.2.7 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. While it was parametrized for adult EFL Japanese readers, it 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 **mcnamara**; François & 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.

$$ML2RI = 164.935 - $$
$$18.792 * WLC - 1.916 * SLW, \qquad (7)$$

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 $[???; -\infty]$.

****

Score of 50 is average dicciuly vf.greenfield 2003:43

formula probably also ok for other l1 backgrounds greenfield 2003:44, at least better than Grade scaled for American classes

(passage frequency: how often does deleted item from cloze test occur in text, not applicable without cloze test),

****

Figure 7: The RFFC workflow.

4.3 Pipeline

The RFFC's pipeline is illustrated in Figure 7. When executing the program, each input file is processed through this pipeline. First, some NLP is used in order to obtain token and sentence boundaries. For this task, implementations of NLTK version 3.2.1 tokenizer are used: sentence segmentation is performed with the nltk.data tokenizer and the German punctuation model, which is loaded from nltk.data "tokenizers/punkt/german.pickle". For tokenization, the nltk.tokenize WordPunctTokenizer is used.

After NLP, the following counts are obtained: number of sentences, number of tokens, number of tokens that are not punctuation marks,[12], number of characters, number of syllables,[13] number of words with one syllable, number of words with one or two syllables, number of words with three or more syllables, number of words with six or more characters, and number of periods and colons.

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, ratio of words to periods and colons, and ratio of sentences to words.

With these features and counts, the twelve readability formulae reported in the previous section are calculated. Features and readability formulae are saved in a document × feature/formula matrix in a csv file.

4.4 Count Robustness

One important issue with readability formulae is that scores are known to differ depending on which implementation is used. To get varying scores for a given text using the same readability formula de-

---

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

[13] 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 **weiss2015** for further information on this issue.
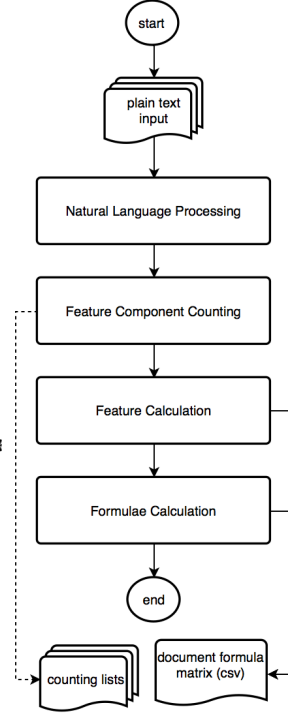
Count	TP	FP	FN	Precision	Recall	F1 Score
tokens	822	8	0	0.990	1	0.995
sentences	61	3	3	0.953	0.953	0.953
words $\geq$ 3 syllables	121	0	1	1	0.992	0.996
words $\leq$ 2 syllables	555	1	0	0.998	1	0.999
words with 1 syllable	324	0	0	1	1	1
words $\geq$ 6 characters	255	0	0	1	1	1
non punctuation tokens	709	0	0	1	1	1

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

pending on the tool that was implementing it, is clearly not desirable, especially, if the source of the difference cannot be investigated easily by inspecting the feature component the formulae are based on, as discussed in subsections 4.1.

Because of this, the RFFC allows for an optional output of the feature component instances that were counted for each document. If the RFFC 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 RFFC 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 RFFC 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 RFFC 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. Accord-

ingly, tokenization handles dates and abbreviations incorrectly, too. Also, the MERLIN corpus consists of letters and, therefore, contains letter heads which challenge the NLP tools, especially with respect to the issues with the processing of dates.

5 Experiment

This section reports on the classification experiments performed to predict the overall Common European Reference Frame (CERF) level attributed to L2 German speakers based on readability formulae applied to German texts written by them. As previously discussed readability scales tend to be notoriously unclear, yet are presumably easy to interpret. Thus, this paper purposely avoids mapping readability scales to CERF levels. The code used to obtain these results may be found in *ml_experiment.py* at ???.

5.1 Data Set

For the proficiency classification experiment the German subsection of the trilingual Merlin corpus was used.[14] While the full Merlin corpus consists of approximately 2,300 texts produced by learners of German, Italian and Czech, in the course of standardized language certifications based on the CERF level scale A1 to C2 the German section has a size of 1,033 texts (Abel et al. 2013; Goullier 2007).

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, coher-
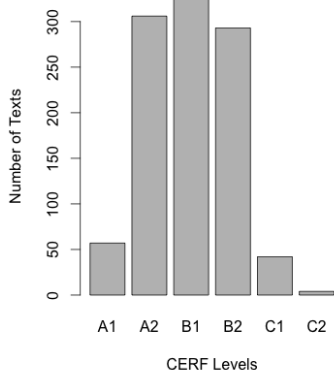
---

[14] http://merlin-platform.eu.

Figure 8: Histogram of grade levels in the German section of the Merlin corpus ($N = 1,033$).

ence 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 8. This issue is addressed in subsection 5.2

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 (2013a,b) and Hancke, Vajjala & Meurers (2012). This makes it especially suited for this task, since these other approaches may be used to evaluate the performance of the readability formulae aside from an uninformed baseline.

5.2 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 each fold approximately balanced. 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 a basic majority baseline and a 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.[15]

5.3 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 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

---

[15]However, results obtained for a less strict significance level of $\alpha \leq 0.05$ are also indicated in Table 4 for sake of completeness.

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. 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 $\alpha \leq 0.05$.

the readability formulae, which is not exceeded by every experiment: none of the formulae except for the CLI perform significantly better on any performance measure. 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 for compared to both baselines, all WSFs, the Fog and the Lix index. It also shows one of the least differences between precision and recall. However, there is no significant difference to either of the Flesch based formulae, the L2 readability formula and 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.

****

Finally, a comparison of the scores scaled to grade levels seems to be in order.

These are the FGS, WSFs, Fog, and CLI.
****

5.4 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 unexpected. 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 formula 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: In this case WSF 3 and 4 should outperform the other two formulae, as they employ less features.

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.

****
"Critics have often cited such discrepancies as indications of the lack of precision of the formulas. ... The most obvious causes of the discrepancies are the different variables used by different formulas and the different criterion scores used in their development. The formulas—like reading tests—simply do not have a common zero point (Klare 1982). ... Finally, the range of scores provided by different formulas remind us that they are not perfect predictors. They provide probability statements or rather estimates of difficulty." Dubay 2004:56

"While the formulas were originally created to help educators select texts for different audiences, writers also use the formula variables to produce texts and transform (re-write) them into simpler versions. The evidence on how effective this is has been mixed. As both the supporters of the formulas and their critics have warned, if you just chop up sentences and use shorter words, the results are not likely to improve comprehension. You have to look at the many other factors that affect reading at the level for which you are writing." dubay 2004:37

Furthermore, it should be noted that this paper did not discuss any sort of mapping from concrete readability scores to CERF scores. However, given the previously discussed criticism on the presumably easy to interpret
****

6 Limitations & Future Work

One issue concerning the experiment is the shortcoming of L2 readability formulae in an L2 proficiency assessment context. They which were represented by the ML2RI, only, a formulae originally designed for Japanese glsefl readers. However, 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 **mcnamara**; François & Fairon (2012). However, regarding the promising performance of the ML2RI it would certainly desirable to enhance the RFFC by

even more L2 formulae.

Also, the questions imposed by the presented experiment could not be answered exhaustively due to the limited scope of this paper. For example, a more details investigation of how learner L1 relates to formula performance would most certainly be of interest. Also, the question of how readability scores assigned to texts differ across formulae has merely been touched. Yet, it is one of the main issues often addressed in the context of readability formula evaluation. On a similar note, an analysis of how the results obtained by the RFFC relate to results obtained by other readability analysis platforms might be interesting at some point.

Last but not least, it would most certainly 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 (2013a,b) and Hancke, Vajjala & Meurers (2012). 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. At least 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.

7 Conclusion

On the one hand, the present paper introduced the RFFC, a python program allowing for an easy to use, yet transparent calculation of various readability formulae. Regarding the high amount of platforms offering diverging scores based on the same readability formula for the same text, yet not giving any indication as to how these differences might be explained, this seemed to be overdue. The RFFC is also publicly available for adjustments and enhancements, serving as a template for further implementations.

On the other hand, this paper also investigated the performance of twelve readability formulae on the task of proficiency assessment in L2 German. Overall, this showed encouraging results, indicating that

readability formulae may in fact also serve as indicators of language proficiency. Interestingly, neither target language nor the relation of target language and reader L1 seemed to be of major influence. In fact, best performance could be obtained for the CLI, which is a very prominent readability formulae, yet neither designed for L2 contexts not for German as target language.

However, these results are by no means to be interpreted such that readability formulae might be suited to provide sufficient feedback for L2 self-study without being conjoined with other methods of performance assessment: As has been established in the course of this paper, readability formulae are easily calculated because they employ superficial features that are indirectly affected by more fine grained features of linguistic complexity. However, this comes with the downside that they are unsuited for exploratory language analyses investigating the interplay of various types of linguistic complexity. In fact, they are completely insensitive to them. Hence, writers, L1 and L2 alike, can neither obtain any explanation on why they received a certain score, nor on how to improve it. For complexity reduction, as desired in most L1 writer contexts, **flesch2016** argues that this is not an issue, as this would happen automatically when decreasing sentence and word length. Be this as it may, in the case of complexity increase, as desired in L2 contexts where complexity is positively correlated with proficiency, this is most certainly not the case: A simple increase in sentence length would most likely fail to generate more proficient writing. For this reason, readability formulae are also unsuited in any contexts where proficiency should actually be graded: Obviously, they are easily tricked in situations where a measured decrease in readability is desired and the formula is known to the writer, for example by simply copying passages.

Yet, this being said, the results of the study might very well be used by learners to perform a quick evaluation on their performance. They should be accompanied by further measures of proficiency, though, and should rather be considered a rough estimation than a robust measure.

References

Abel, Andrea et al. (2013). *merlin: A Trilingual Learner Corpus illustrating European Reference Levels*. LRC 2013. Bergen, Norway.

Amstad, T. (1978). "Wie verständlich sind unsere Zeitungen?" PhD thesis. University of Zürich.

Björnsson, C. H. (1983). "Readability of Newspapers in 11 Languages". In: *Reading Research Quarterly* 18.4, pp. 480–497.

Brettschneider, Frank (2012). *FAQs im Klartext-Check: Statt Antworten erhalten Kunden oft nur Kauderwelsch*. Tech. rep. Press and publicity work of the University of Hohenheim.

Brettschneider, Frank (2013). *Kauderwelsch statt Klartext: Kundenkommunikation von Telekommunikations-Unternehmen ist oft unverständlich*. Tech. rep. Press and publicity work of the University of Hohenheim.

Brettschneider, Frank (2014). *CEO-Reden im Vergleich: CEO-Reden im Vergleich: Am verständlichsten ist BMW-Chef Norbert Reithofer*. Tech. rep. Press and publicity work of the University of Hohenheim.

Brettschneider, Frank (2016). *Langzeitstudie Wahlprogramme: Parteien bemühen sich wieder zunehmend um Verständlichkeit*. Tech. rep. Press and publicity work of the University of Hohenheim.

Brown, James Dean (1997). "An EFL readability index". In: *University of Hawai'i Working Papers in English as a Second Language* 15.2, pp. 85–119.

Chall, J. S. (1958). *Readability: An appraisal of research and application*. Ohio State University Press.

Coleman, Meri & T. L. Liau (1975). "A Computer Readability Formula Designed for Machine Scoring". In: *Journal of Applied Psychology* 60.2, pp. 283–284.

Crossley, Scott A., Jerry Greenfield & Danielle S. McNamara (2008). "Assessing Text Readability Using Cognitively Based Indices". In: *Tesol Quarterly* 42.3, pp. 475–493.

DuBay, William H. (2004). "The Principles of Readability". In: *Online Submission*.

Esfahani, B Janghorban et al. (2014). "Systematic analysis of the readability of patient information on the websites of clinics for plastic surgery". In: *Handchirurgie, Mikrochirurgie, plastische Chirurgie: Organ der Deutschsprachigen Arbeitsgemeinschaft fur Handchirurgie: Organ der Deutschsprachigen Arbeitsgemeinschaft fur Mikrochirurgie der Peripheren Nerven und Gefasse* 46.6, pp. 369–374.

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

François, Thomas & 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*. Association for Computational Linguistics. Jeju Island, Korea, pp. 466–477.

Goullier, Francis (2007). *The Common European Framework of Reference for Languages (CERF) and the developement of language policies: challenges and responsibilities*. Tech. rep. Intergovernmental Language Policy Forum.

Greenfield, Jerry (1999). "Classic readability formulas in an EFL context: Are they valid for Japanese speakers?" PhD thesis. Temple University.

Greenfield, Jerry (2003). "The Miyazaki EFL Readability Index". In: *Comparative Culture* 9, pp. 41–49.

Greenfield, Jerry (2004). "Readability Formulas For EFL". In: *Japan Association for Language Teaching* 26.1, pp. 5–24.

Gunning, Robert (1968). *The technique of clear The technique of clear writing*. New York: McGraw-Hill.

Hamsik, M. J. (1984). "Reading, readability, and the ESL reader". PhD thesis. University of South Florida.

Hancke, Julia (2013a). "Automatic Prediction of CERF Proficiency Levels Based on Linguistic Features of Learner Language". MA thesis. Eberhard Karls Universität Tübingen.

Hancke, Julia (2013b). "Exploring of CERF classification for German based on rich linguistic modeling". In: *Learner Corpus Research Conference*, pp. 54–56.

Hancke, Julia, Sowmya Vajjala & Detmar Meurers (2012). "Readability Classification for German using lexical, syntactic and morphological features". In: *Proceedings of COLING*. Mumbai, pp. 1063–1080.

Klare, G. R. (1963). *The measurement of readability*. Iowa State University Press.

Leichter, Steven B et al. (1981). "Readability of self-care instructional pamphlets for diabetic patients". In: *Diabetes Care* 4.6, pp. 627–630.

Leong, Elain K. F., Michael T. Ewing & Leyland F. Pitt (2002). "E-comprehension. Evaluating B2B websites using readability formulae". In: *Industrial Marketing Management* 31, pp. 125–131.

Luers, J.-C. et al. (2013). "Lesbarkeit von medizinischen Texten im Internetangebot deutscher HNO-Universitätskliniken". In: *HNO* 61.8, pp. 648–654.

Mailloux, Steven L et al. (1995). "How Reliable is Computerized Assessment of Readability?" In: *Computers in nursing* 13, pp. 221–221.

Smith, E. A. & R. J. Senter (1967). *Automated Readability Index*. Tech. rep. DTIC Document.

von der Brück, Tim (2007). "A Semantically Oriented Readability Checker for German". In: *Proceedings of the 3rd Language and Technology Conference*. Poznań, Poland, pp. 270–274.

von der Brück, Tim (2008). "A Readability Checker with Supervised Learning Using Deep Indicators". In: *Informatica* 32, pp. 429–435.