sciendo
DOI 10.2478/jazcas-2021-0044

USING A PARALLEL CORPUS TO ADAPT THE FLESCH READING EASE FORMULA TO CZECH

KLÁRA BENDOVÁ
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

BENDOVÁ, Klára: Using a parallel corpus to adapt the Flesch Reading Ease formula to Czech. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 477 – 487.

Abstract: Text readability metrics assess how much effort a reader must put into comprehending a given text. They are, e.g., used to choose appropriate readings for different student proficiency levels, or to make sure that crucial information is efficiently conveyed (e.g., in an emergency). Flesch Reading Ease is such a globally used formula that it is even integrated into the MS Word Processor. However, its constants are language-dependent. The original formula was created for English. So far it has been adapted to several European languages, Bangla, and Hindi. This paper describes the Czech adaptation, with the language-dependent constants optimized by a machine-learning algorithm working on parallel corpora of Czech and English, Russian, Italian, and French, respectively.
Keywords: complexity, parallel corpus, Czech, Flesch Reading Ease, machine learning

1 INTRODUCTION

This study describes a machine learning-based approach to adapting the widely known Flesch Reading Ease [1] formula to Czech, based on a parallel corpus [2].

A written text is always a message conveyed by the author to the recipient without real-time interaction. Therefore, the author must assess the intended reader well, regarding their knowledge of the topic and contexts, but also their reading comprehension skills. This is immensely important, whenever lives and health, security, democracy, or property are at stake.

This is where the concept of readability comes into play. DuBay [3, p. 6] has summarized its most prominent definitions: „readability is the ease of reading created by the choice of content, style, design, and organization that fit the prior knowledge, reading skill, interest, and motivation of the audience". Particularly in the English-speaking community, quantitative assessment of readability has been worked on since the early 20th century. The 1980's have already seen over 200 different readability formulas, with over a thousand studies attesting to their strong theoretical and statistical validity [3].

One of the most common readability formulas is Flesch Reading Ease [1], which is even implemented in the MS Word editor. On a scale from 0 to 100, it measures the „ease“ of the text, using general features such as average length of sentences in words and average length of words in syllables, and a few constants. However, these constants are language-dependent. Šlerka and Smolík [4] found out in their pilot experiment that the Flesch Reading Ease was associated with the intuitively perceived linguistic complexity of different text genres even in Czech. However, the scores would not fall between 0 and 100. Due to inflections and the absence of articles, both making the average Czech word longer than English, any natural Czech text would score as difficult. Even common newspaper texts reach negative values, beyond the extreme difficulty end of the English scale.

This study will (1) introduce a selection of tools assessing diverse complexity features of Czech and other, mainly Slavic, languages; (2) describe the Flesh Reading Ease formula and its existing language adaptations; (3) describe the data and its pre-processing to derive the Czech parameters for Flesch Reading Ease; (4) describe the experiment; (5) report and interpret its results. Its goal is to offer a Czech-tailored replacement for the original English-based Flesch Reading Ease for the assessment of the readability of Czech texts.

2 RELATED WORK

2.1 Tools

There are numerous online tools to assess readability of English texts by diverse formulas. Nevertheless, this section will only list tools that were immediately relevant for this study. These are mostly tools tailored to Czech or Polish, and a multilingual tool that is still in development.

One of the most inspiring tools is EVALD [5], which primarily assesses text cohesion and coherence in pupils’ essays, predicting the grade given by teachers. It is partly based on international readability formulas Flesh Reading Ease [1], Flesch-Kincaid Grade Level Formula [6], Coleman-Liau index [7], SMOG index [8], but none of them has been adapted to Czech. Apart from the cohesion/coherence assessment, EVALD has also been trained on Czech texts written by foreigners to guess the CEFR [9] proficiency level of their authors [10].

Another text assessment tool for Czech is QuitaUp [11], which mainly captures stylometric characteristics such as TTR, h-point, entropy, or word distance.

However, neither is a dedicated readability tool like e.g., the Polish Jasnopis [12], which combines statistical features with empirically measured reading comprehension.

Eventually, a multilingual readability assessment platform is in development (Common Text Analysis Platform – CTAP) [13]. It aggregates 600 textual features ranging from syllable count to lexical sophistication tailored to the languages currently represented: English and German. Other languages being worked on are Italian, French, Portuguese, Greek, and Czech.

3 FLESCH READING EASE

3.1 The original English formula

Flesch Reading Ease, presented by Rudolf Flesch [1], is defined as follows:

FRE = 206.835 – 84.6 wl – 1,015 sl[1],

where FRE = Flesch Reading Ease
wl = average word length in syllables
sl = average sentence length in words.

Henceforth I will refer to exact values as coefficients, while calling the variable formula elements parameters.

The results of Flesch Reading Ease virtually always fit within the range of 0 – 100. The higher the score, the higher the "ease," that is, the more its complexity decreases, and the lower education is expected in the reader to be well equipped to comprehend the text.

Flesch interprets these results in the book The Art of Readable Writing [14] as follows:

Reading Ease Score	Style Description	Estimated Reading Grade
0 to 30:	Very Difficult	College graduate
30 to 40:	Difficult	13th to 16th grade
50 to 60:	Fairly Difficult	10th to 12th grade
60 to 70:	Standard	8th and 9th grade
70 to 80:	Fairly Easy	7th grade
80 to 90:	Easy	6th grade
90 to 100:	Very Easy	5th grade

Tab. 1. Flesch Reading Ease Index interpretation

When computing this formula, Flesch was drawing on a formula he had invented in 1943. He skipped affix counts since they had proved troublesome to count for the formula users. Instead, he transformed this feature into syllable count, which he considered more mechanical and thus less error-prone [15]. However, Flesch used the omitted counts to determine the coefficients.

[1] I am quoting the paper *A New Readability Yardstick* [1] from DuBay's compilation of readability studies *Unlocking Language: The Classic Readability Studies* [3], where the decimal separator is misplaced (FRE = 206.835 – 84.6 wl – 1.015 sl), whereas the formula correctly reads FRE = 206.835 – 84.6 wl – 1,015 sl.

3.2 Language mutations of Flesch Reading Ease

Even formulas that use very generic features are as heavily language-dependent, as languages differ with respect to their phonological, morphological, and syntactic features. Individual languages need individual formulas. This also even applies to Flesch Reading Ease. Guryanov et al. [16] interpret its parameters as follows: WL (word length as the ratio between total syllables and total tokens) renders the information load of the text; short words make the text less informative than long words. SL (sentence length as the ratio between total words and total sentences) reflects cohesion; that is, cohesion decreases with the sentence length. This difference is language-dependent. I. V. Oborneva [17] observed that an average English word has 2.97 syllables, while an average Russian word has 3.29 syllables. This necessarily affects the coefficients; the more so if the results are supposed to span the same scale and be cross-linguistically comparable.

Currently there are formulas for Italian, French, Spanish [18], German [19], Russian [17], and Danish, as well as for Bangla and Hindi [20]. Garais [18] also mentions a Japanese formula, but the source is not sufficiently quoted.

The formulas were designed at different times, with different methods available then. The more recent formulas draw on machine-learning algorithms run over large data, including parallel corpora, while older formulas are based on sophisticated calculations.

For French, the first Flesch versions were calculated in 1958 [21] and 1963 [22], to be replaced by a third version [23], which is still in use [24]. Despite extensive research, unfortunately limited to the English-written literature, I have failed to find this current version for French and had to resort to the 1958 version [21].

FRE(French) = 207 – 1.015(total words/total sentences) – 73.6(total syllables/ total words)

The first version of the Russian formula was designed by Matkovskij in the 1970's. Matkovskij grounded his formula in the fact that Russian words have, on average, more syllables than English words and, therefore, he replaced one of the parameters with the number of tokens that have more than three syllables (Matkovskij, 1976 in [25]).

FRE(Russian_mod)= 0.62(total words/total sentences) + 0.123 X3 + 0,051

where X3 = the percentage of tokens with more than three syllables.

A more recent Russian version came from Oborneva in 2006. As already mentioned above, Oborneva based her calculations on the difference in number of syllables in Russian and English words [16], drawing on *Slovar russkogo yazyka pod redaktsyey Ozhegova* (39174 words) and *Muller English-Russian dictionary* (41977 words). In addition, she analyzed six million words of parallel Russian-English literary texts [16], her work resulting in the following formula [17]:

FRE(Russian) = 206.835 – 1.3(total words/total sentences) – 60.1(total syllables/total words).

4 DATA

The experiment is based on a cross-lingual comparison of parallel texts; therefore I used data from the InterCorp parallel corpus ([26], [2]). This corpus has Czech as the pivot language: all texts have a Czech version, which is manually sentence-aligned with at least one different language. Foreign languages are never directly aligned with each other, but through Czech. The Czech texts are both original texts, as well as translations. Among foreign texts, originals or translations from Czech were preferred during the acquisition, but translations from other languages are present as well. The corpus primarily comprises fiction, but also non-fiction and legal texts from the multilingual official production of the EU bodies. Tab. 2 shows the distribution of selected languages in InterCorp.

Language	Czech	English	Russian	French	Italian
Total of texts	586	348	128	233	136
Total of sentences	3 719 974	2 364 684	855 584	1 160 089	992 008
Total of tokens	43 446 132	33 190 659	9 449 802	18 921 311	14 466 499

Tab. 2. Distribution of the data used

5 METHOD

This section describes the actual experiment. Its goal was adaptation of Flesch Reading Ease to Czech and assessment of its validity by comparison with formulas for other languages.

The parameters of the Flesch Reading Ease formula are counts of words, syllables, and sentences. The InterCorp data came as XML files with tokenization and sentence splitting. I tested the sentence splitting with UDPipe [27], with no resulting corrections. I used the same method to count words and sentences in all languages.

On the other hand, syllable counting required individual language-specific scripts, since the phonotactic rules, as well as phoneme distributions, are language-specific. My syllable-counting scripts were based on a syllable-counting script by David Lukeš from the Institute of the Czech National Corpus, which considers the pitch (a vowel, diphthong, or a syllabic consonant), rather than syllable boundaries. Another option was using the PyHyphen library[2], as done for instance in Jasnopis [12], but my rule-based scripts were giving better results in manual sample checks. However, both approaches had problems counting syllables in French. The complexity of French

[2] Available at: https://pypi.org/project/PyHyphen/.

syllable counting can explain worse experiment results for French. When processing Russian, I considered only vowels to form syllables, drawing on [25].

Figure 1 shows the curves of the language specific FRE scores on parallel texts from InterCorp. Considering the English curve the reference, the Russian FLE fits it far better than the French and Italian. This implies that the Italian and French formulas, at least in my implementation, are less suitable to train the Czech formula adaptation than Russian, to achieve the best possible fit to English.

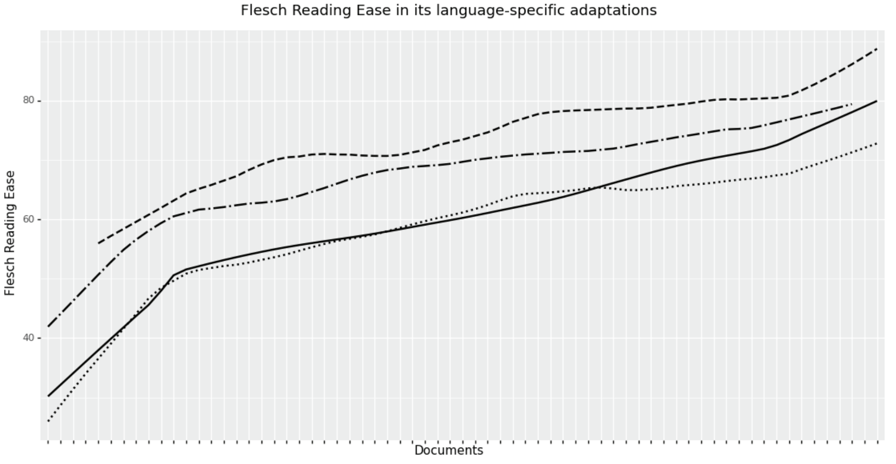

Fig. 1. Flesch Reading Ease in its language-specific adaptations. Solid line _ for English, dotted line **...** for Russian, dashdot line for French and dashdash line for Italian

To quantify the deviations seen in the plot in Fig. 1, I computed the RMSE (Root Mean Squared Error, a standard deviation evaluation in machine learning) of each language-specific FRE to the English FRE on English (Tab. 3). The French and Italian RMSE are indeed substantially higher, as expected based on Fig. 1.

	RMSE
English	–
Russian	5.100
French	10.518
Italian	12.991

Tab. 3. Root mean squared error for every language-dependent FRE used on Czech documents compared to English FRE used on the corresponding documents in English

The scatterplot in Figure 2 shows the FRE curves of Czech documents computed with the individual language-specific formulas; that is, the English, French, Italian, and Russian FRE for each Czech document, distinguished by the point shape. The solid line shows the *English* FRE of the corresponding *English* versions of the Czech documents, as a reference of accuracy. The documents (on the X-axis) are ordered according to the English FRE of their English versions. There is an observable difference between the reference English curve and the curve representing the English FRE on the Czech documents. The original English FRE formula presents the Czech texts almost twenty points lower, which says that the Czech texts are two reading proficiency levels more difficult than their English versions.

Without fitting FRE to Czech, the best language-dependent formula to use would be the Russian one. It can certainly lie in the closeness of these two languages, but it can also be attributed to the worse fit of the French and Italian formulas to the English original (cf. Fig. 1).

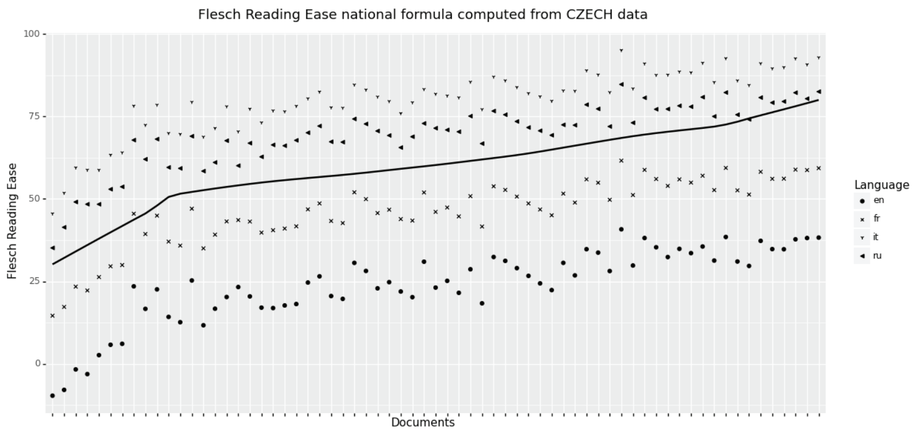

Fig. 2. Language-specific Flesch Reading Ease formula computed on Czech data. Solid line _ for English as the reference value

To find the optimal parameters for Flesch Reading Ease, I used the optimize. curve_fit algorithm from the SciPy library [28] with Russian and English separately. I neglected French and Italian due to their substantially worse fit to English. On input, the algorithm got FRE values of the individual Czech documents computed with the corresponding formula for the reference language. The algorithm compared these values with the values of the corresponding documents in the corresponding foreign language. The outcome was two different FRE functions for Czech.

I repeated the experiment with documents chunked into 100-sentence batches to increase the number of observations. The English and Russian inputs increased from 348 observations to 19,722 and 128 to 6,138 observations, respectively. However, this has not affected the best fit made on Russian texts, shown in Tab. 4. The best result was obtained using whole Russian texts as reference with RMSE 3.748 on test data.

Text types (number)	FRE for CZECH	RMSE test data
EN texts (347)	$206.835 - 1.424(\text{tot words}/\text{tot sentences}) - 63.920(\text{tot syllables}/\text{tot words})$	6.039
EN parts (19 722)	$206.835 - 1.672(\text{tot words}/\text{tot sentences}) - 62.182(\text{tot syllables}/\text{tot words})$	4.639
RU texts (127)	**$206.835 - 1.388(\text{tot words}/\text{tot sentences}) - 65.090(\text{tot syllables}/\text{tot words})$**	**3.748**
RU parts (6 138)	$206.835 - 1.514(\text{tot words}/\text{tot sentences}) - 60.096(\text{tot syllables}/\text{tot words})$	4.363

Tab. 4. Version of Flesch Reading Ease for Czech language and the RMSE computed for test data

For the final evaluations, I merged the train and test data for English, Czech, and Russian, respectively, and computed the RMSE between the Czech FRE and English FRE, as well as the RMSE between the Russian and the English FRE. The Czech-English RMSE is 5.067, which is better than RMSE for Russian and English with 5.100 (Tab. 5).

	RMSE
English	–
Russian	5.100
French	10.518
Italian	12.991
Czech	**5.067**

Tab. 5. Root mean squared error for every language compared to English

Figure 3 confirms that the Czech language specific FRE on Czech texts is closest to the English FRE on English texts from all available language specific FREs on their languages.

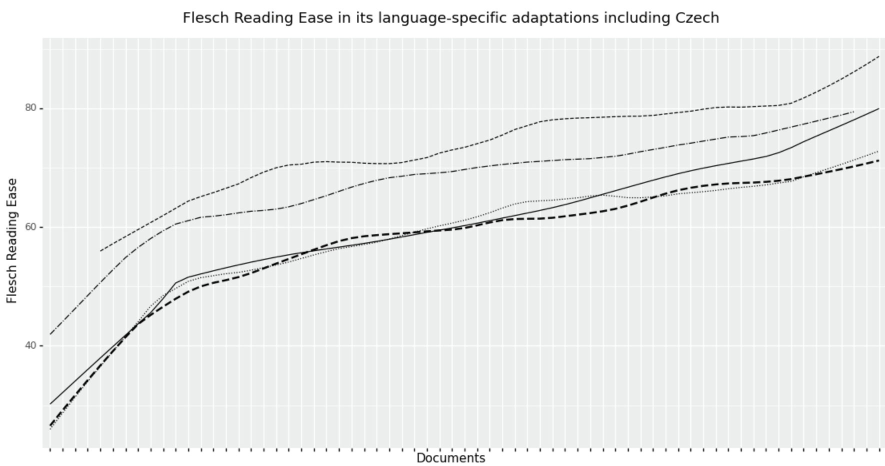

Fig. 3. Flesch Reading Ease in its language-specific adaptations including Czech. Original lines from Fig. 1 are thin, while the one for Czech formula -- is thick

6 DISCUSSION

Although the parallel corpus is relatively small and it is not balanced, the Czech formula superseded the (obsolete) French and Italian formulas. The relatively poor fit of the French and Italian formulas to English compared to Russian in this exercise can be blamed on possible conceptual errors in my syllable-counting scripts, while the fit on Russian was so much better because substantial syllable-conceptualization differences are very unlikely in this language pair. I have reached the maximum possible fit given the available data and language-dependent formulas.

This work is part of a larger project. The Czech adaptations of this and other readability formulas and features are to be implemented in CTAP [13]. The script is freely available at GitHub [29].

This entire approach naturally draws on the assumption that translations have the same readability as originals. Good translations are supposed to be semantically and stylistically faithful, as well as idiomatic. Given that InterCorp comprises mainly professionally published fiction and official multilingual documents, the translation quality is maintained.

The statistics (word and syllable counts) on which the current FRE is based are seemingly primitive, but Flesch himself proved them to strongly correlate with much more sophisticated statistics he had used earlier. In the original formula versions, Flesch made use of the contemporary psychological and pedagogical knowledge and found text features to reflect how “conversational”, “personal”, and “interesting” a text passage be, considering also text cohesion by counting pronouns, personal names, and nouns referring to humans. Besides, he accounted for the conceptual complexity (abstraction) by counts of lexical derivatives [15, p. 101]

These units are so essential for the content that they do not leave much room for deviation between languages. This suggests that, although their counts will be different in translation pairs (e.g., pronouns between a pro-drop and non-pro-drop language), their distributions within each language will be similar. The dissimilarity creates the documented error margins of the individual formula adaptations.

ACKNOWLEDGEMENTS

This work was supported by the Czech Science Foundation grant 19-19191S: Linguistic Factors of Readability in Czech Administrative and Educational Texts. The work described herein has also been using data/tools/services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

References

[1] Flesch, R. (1948). A New Readability Yardstick. Journal of Applied Psychology, 32, pages 221–233.

[2] Rosen, A. (2016). InterCorp – a look behind the façade of a parallel corpus. In Polskojęzyczne Korpusy Równoległe Polish-Language Parallel Corpora, pages 21–40, Instytut Lingwistyki Stosowanej, Warszawa.

[3] DuBay, W. (2007). Smart Language. Readers, Readability, and the Grading of Text. Impact Information, Costa Mesa, California.

[4] Šlerka, J., and Smolík, F. (2010). Automatická měřítka čitelnosti pro česky psané texty. Studie z Aplikované Lingvistiky, 1, pages 33–44.

[5] Novák, M., Mírovský, J., Rysová, K., Rysová, M., and Hajičová, E. (2019). EVALD 4.0 – Evaluator of Discourse. Accesible at: http://hdl.handle.net/11234/1-3065.

[6] Kincaid, J. P., Fishburne, R. P., Rogers, R. L., Chissom, B. S., and BRANCH, N.T.T.C.M.T.R. (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula). for Navy Enlisted Personnel. Defense Technical Information Center. Accessible at: https://books.google.cz/books?id=7Z7ENwAACAAJ.

[7] Coleman, M., and Liau, T. L. (1975). A computer readability formula designed for machine scoring. Journal of Applied Psychology, 60, pages 283–284.

[8] McLaughlin, G. H. (1969). SMOG grading – a new readability formula. Journal of Reading, 22, pages 639–646.

[9] Council of Europe. (2018). Common European Framework of Reference for Languages: Learning, Teaching, Asessment. Companion volume. Council of Europe Publishing, Strasbourg. Accesible at: https://www.coe.int/lang-cefr.

[10] Rysová, K., Rysová, M., Mírovský, J., and Novák, M. (2017). Introducing EVALD – Software Applications for Automatic Evaluation of Discourse in Czech. RANLP Proceedings, Bulgaria, pages 634–641.

[11] Cvrček, V., Čech, R., and Kubát, M. (2020). QuitaUp. Czech National Corpus and University of Ostrava. Accesible at: https://www.korpus.cz/quitaup/.

[12] Dębowski, Ł., Broda, B., Nitoń, B., and Charzyńska, E. (2015). Jasnopis – A Program to Compute Readability of Texts in Polish Based on Psycholinguistic Research. Natural Language Processing and Cognitive Science, 2015 Libreria Editrice Cafoscarina, Venezia, Italy, pages 51–61.

[13] Chen, X., and Meurers, D. (2016). CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. Apollo – University of Cambridge Repository. Accesible at: https://www.repository.cam.ac.uk/handle/1810/292470.

[14] Flesch, R. (1974). The art of readable writing. 2nd ed. Harper, New York.

[15] DuBay, W. H. (2008). Unlocking Language: Classic Readability Studies. IEEE Transactions on Professional Communication, 51.

[16] Guryanov, I., Yarmakeev, I., Kiselnikov, A., and Harkova, I. (2017). Text Complexity: Periods of Study in Russian Linguistics. Revista Publicando, 4, pages 616–625.

[17] Oborneva, I. V. (2006). Mathematical model for evaluation of didactic texts. Proc of Moscow State Pedag Univ, 4, pages 141–147.

[18] Garais, E.-G. (2011). Web Applications Readability. Romanian Economic Business Review, 5, pages 117–121.

[19] Amstad, T. (1978). Wie verständlich sind unsere Zeitungen? Studenten-Schreib-Service.

[20] Sinha, M., Sharma, S., Dasgupta, T., and Anupam, B. (2012). New Readability Measures for Bangla and Hindi Texts.

[21] Kandel, L., and Moles, A. (1958). Application de l'indice de flesch à la langue française. Cahiers Etudes de Radio-Télévision, 19, pages 253–274.

[22] De Landsheere, G. (1963). Pour une application des tests de lisibilité de Flesch à la langue française. Le Travail Humain, pages 141–154.

[23] Henry, G. (1975). Comment mesurer la lisibilité. Labor, Brussels, Belgium.

[24] François, T., and Fairon, C. (2012). An AI readability formula for French as a foreign language, 477 p.

[25] Solnyshkina, M., Ivanov, V., and Solovyev, V. (2018). Readability Formula for Russian Texts: A Modified Version: 17th Mexican International Conference on Artificial Intelligence, MICAI 2018, Proceedings, Part II, pages 132–145.

[26] Čermák, F., and Rosen, A. (2012). The Case of InterCorp, a multilingual parallel corpus. International Journal of Corpus Linguistics, 13, pages 411–427.

[27] Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics, Brussels, Belgium, pages 197–207.

[28] SciPy 1.0 Contributors, Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T. et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods, 17, pages 261–272.

[29] https://github.com/vanickovak/ReadabilityFormula.