1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI2020), pages 77–84
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

A multi-lingual and cross-domain analysis of features for text simplification

Regina Stodden and Laura Kallmeyer
Heinrich Heine University
Düsseldorf, Germany
{stodden, kallmeyer}phil.hhu.de

Abstract

In text simplification and readability research, several features have been proposed to estimate or simplify a complex text, e.g., readability scores, sentence length, or proportion of POS tags. These features are however mainly developed for English. In this paper, we investigate their relevance for Czech, German, English, Spanish, and Italian text simplification corpora. Our multi-lingual and multi-domain corpus analysis shows that the relevance of different features for text simplification is different per corpora, language, and domain. For example, the relevance of the lexical complexity is different across all languages, the BLEU score across all domains, and 14 features within the web domain corpora. Overall, the negative statistical tests regarding the other features across and within domains and languages lead to the assumption that text simplification models may be transferable between different domains or different languages.

Keywords: text simplification, corpus study, multi-lingual, multi-domain

1. Introduction

In research regarding readability and text simplification, several features are mentioned which identify easy-to-read sentences or help to transform complex to simplified texts. However, features such as readability metrics are highly criticized because they only consider surface characteristics, e.g., word and sentence length, ignore other relevant factors, such as infrequent words (Collins-Thompson, 2014), and are optimized only for English. Therefore, Collins-Thompson (2014) proposes more sophisticated features, e.g., parse tree height or word frequency, which might be applicable to non-English-languages too.

Similar to the research in text readability, most text simplification research is concerned with English, with some exceptions, e.g., Italian (Brunato et al., 2016) or Czech (Barančíková and Bojar, 2019), or multi-lingual approaches, e.g., Scarton et al. (2017). Text simplification or readability measurement models with the same feature set for all corpora have been shown to perform well on cross-lingual (Scarton et al., 2017), multi-lingual (Yimam et al., 2017), and cross-domain (Gasperin et al., 2009) corpora. However, due to language or domain characteristics, distinct features, e.g., parse tree height, proportion of added lemmas, or usage of passive voice, might be more or less relevant during the simplification process and also during its evaluation. So far, it has not been investigated whether the relevance of distinct text simplification features differs across languages and domains. We therefore address the following research questions (RQ) in this paper:

1. Do complex texts and its simplified version differ significantly regarding linguistic features? Can language-independent linguistic features explain at least partially the simplification process?

2. Is the simplification process consistent between corpora across and within domains?

3. Is the simplification process consistent between corpora within and across languages?

Concretely, we analyze the relevance of features named in readability and text simplification research on aligned sentence simplification pairs in five languages, i.e., Czech, German, English, Spanish, and Italian, and in three domains, i.e., web data, Wikipedia articles, and news articles. This automated multi-lingual text simplification corpus analysis is implemented based on the analysis proposed in Martin et al. (2018). For re-use on other corpora, our code is available on github[1].

The paper is structured as follows: Section 2 gives an overview of related work, the next section describes our methods for addressing the above mentioned research questions, including corpora, features, and evaluation methods. Section 4 discusses our results, and Section 5 concludes.

2. Related Works

Several studies of text readability/simplification analyze or compare texts or sentence pairs with different complexity levels, e.g., Collins-Thompson (2014) or Kauchak et al. (2014) in English, Hancke et al. (2012) in German, Gasperin et al. (2009) or Aluisio et al. (2010) in Portuguese, Pilán and Volodina (2018) in Swedish, and Scarton et al. (2017) in English, Italian, and Spanish. However, in contrast to the paper in hand, they focus on building either complexity level assessment models using and comparing grouped features sets or on the theoretical justification of these features (Collins-Thompson, 2014) rather than on a comparison of the relevance and statistical significance of the distinct features (see RQ1). Most of the text level features proposed in these studies, e.g., parse tree height, passive voice, length of verb phrases, are also considered in our work. Unfortunately, we could not include discourse-level features, e.g., coherence, idea density, or logical argumentation, because of the lack of alignments at that level.

In the context of text simplification, several related corpus studies exist either to analyze the quality of a new corpus,

[1] https://github.com/rstodden/TS_corpora_analysis

e.g., (Xu et al., 2015) or Scarton et al. (2018), or to build an evaluation metric, e.g., Martin et al. (2018). Martin et al. (2018) implemented several features regarding English text simplification and test whether they correlate with human judgments in order to build an evaluation metric which does not require gold simplifications. Their work is the most similar to ours, but in comparison to them, we will analyze simplification features from another perspective: Instead of comparing with human judgments, we will evaluate the features at their simplification level, language, and domain. The analysis proposed here is based on their implementation, but it extends it with more features and enables the analysis of other languages than English.

Gasperin et al. (2009) built a classifier that predicts whether a sentence needs to be split in the context of Portuguese text simplification. Their basic feature set, including, e.g., word length, sentence length, and number of clauses, achieved good results on the news-article domain (F-score of 73.40), the science articles domain (72.50) but performs best cross-domain (77.68). We use similar features but analyze them separately and evaluate them regarding other domains, i.e., web data and Wikipedia (see RQ2).

The topic of multi-lingual text simplification is also related to this paper. For complex word identification, a sub-task of text simplification, a data set in German, English, and Spanish exists (Yimam et al., 2017). On this data set, Finnimore et al. (2019) tested language-independent features as to whether they generalize in a cross-lingual setting. Their ablation tests identified the number of syllables, number of tokens, ratio of punctuation, and word probability as the best performing features. In contrast, Scarton et al. (2017) focus on syntactical multi-lingual simplification. They proposed a multi-lingual classifier for deciding whether a sentence needs to be simplified or not for English, Italian, and Spanish, using the same features for all languages. For each language, the system achieved an F1-score of roughly 61% using the same feature set. In our study, we investigate whether their findings also hold for both syntactic and lexical simplifications and not only one of them (see RQ3).

3. Method

In order to compare text simplification corpora in different languages and domains, we have chosen eight corpora in five languages and three domains (see Section 3.1). For the analysis, we use in sum 104 language-independent features (see Section 3.2). In order to analyze relevance of the features per corpus, language, and domain, we conduct several statistical tests (see Section 3.3).

3.1. Data

Most text simplification research focuses on English, but also research in other languages exist, e.g., Bulgarian, French, Danish, Japanese, Korean. However, due to limited access, now-defunct links, non-parallel-versions, or a missing statement regarding availability, we focus on the following four non-English text simplification corpora:

- German (DE) web data corpus (Klaper et al., 2013),
- Spanish (ES) news corpus Newsela (Xu et al., 2015)[2],
- Czech (CS) newspaper corpus COSTRA (Barančíková and Bojar, 2019)[3], and
- Italian (IT) web data corpus PaCCSS (Brunato et al., 2016)[4].

In contrast, several freely available corpora for English text simplification exist. We decided to use the following four:

- TurkCorpus (Xu et al., 2016)[5],
- QATS corpus (Štajner et al., 2016)[6], and
- two current used versions of the Newsela corpus (Xu et al., 2015)[7].

The first version of Newsela (2015-03-02) (Xu et al., 2015) is already sentence-wise aligned whereas the second version (2016-01-29) is not aligned. Therefore, the alignment is computed on all adjacent simplification levels (e.g., 0-1, 1-2, .., 4-5) with the alignment algorithm MASSAlign proposed in Paetzold et al. (2017)[8] using a similarity value α of 0.2 for the paragraph as well as for the sentence aligner.

In addition to the language variation, the corpora chosen for this purpose differ in their domains, i.e., newspaper articles, web data, and Wikipedia data. An overview, including the license, domain, size, and alignment type of the corpora, is provided in Table 1.

As illustrated in Table 1, the corpora largely differ in their size of pairs (CS-Costra: 293, EN-Newsela-15: 141,582) as well as in the distribution of simplification transformations (see Table 1), e.g., 15% of only syntactic simplifications in EN-QATS but only 0.03% in EN-Newsela-15.

3.2. Features

For the analysis, overall, 104 language-independent features are measured per corpus, domain, or language. 43 features, further called *single features*, are measured per item in the complex-simplified pair. For the domain and language comparison, the difference of each of the same 43 features between the complex and simplified text is measured, further called *difference feature*s. The remaining 18 features, *paired features*, describe respectively one feature per complex-simplified pair. The implementation of the features is in Python 3 and is based on the code provided by Martin et al. (2018). In contrast to them, we are offering the usage of SpaCy[9] and Stanza[10] instead of NLTK for pre-processing. In comparison to SpaCy, Stanza is slower but has a higher accuracy and supports more languages. In the following, the results using SpaCy are presented.

[2] https://newsela.com/data/

[3] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3123

[4] http://www.italianlp.it/resources/paccss-it-parallel-corpus-of-complex-simple-sentences-for-italian/

[5] https://github.com/cocoxu/simplification

[6] http://qats2016.github.io/shared.html

[7] https://newsela.com/data/

[8] The code of the tool is originally published in Python 2. The tool was used in Python 3 following the code published at https://github.com/samuelstevens/massalign.

[9] https://spacy.io/

[10] https://stanfordnlp.github.io/stanza/

		Domain	Size	License	Sentence length		Word length		S	L	L&S	-L&-S	I
					comp	simp	comp	simp					
CS-COSTRA		News Headlines	293	CC BY 4.0	11.47	9.65	5.43	5.29	1.02	58.02	40.61	0.34	0
DE-Klaper		Web Data	1,888	Available Upon Request	12.79	12.45	6.98	6.34	1.06	20.71	57.20	21.03	20.60
EN-Newsela_15		News Articles	141,582	Scientific Usage	26.27	17.25	5.32	5.15	0.03	37.21	62.75	0.01	0
EN-Newsela_16	0-1	News Articles	69,185	Scientific Usage	25.50	24.63	4.99	4.96	2.87	24.60	27.73	44.81	41.54
	1-2		76,533		21.35	20.41	4.99	4.94	3.35	26.67	29.48	40.50	36.85
	2-3		69,229		18.18	16.94	4.93	4.84	3.78	29.53	34.60	32.10	28.80
	3-4		61,383		15.17	14.05	4.84	4.76	4.13	28.60	34.82	32.45	29.15
	4-5		966		12.36	10.87	4.64	4.52	4.14	32.40	39.75	23.71	21.12
EN-QATS		Wikipedia + Encyclopedia	505	Free Usage	28.20	23.53	5.35	5.32	15.45	29.70	30.50	24.36	18.61
EN-Turk		Wikipedia	18,872	GNU General Public License	22.35	21.37	5.38	5.20	3.50	47.50	28.05	20.96	15.96
ES-Newsela	0-1	News Articles	7,529	Scientific Usage	31.31	28.96	5.28	5.25	2.36	38.23	37.06	22.35	19.83
	1-2		8,235		25.87	23.77	5.26	5.22	3.05	33.71	34.73	28.51	24.53
	2-3		6,783		21.36	18.81	5.21	5.13	3.07	35.03	40.31	21.60	19.08
	3-4		5,707		16.35	14.43	5.13	5.07	3.31	33.70	39.41	23.59	20.90
	4-5		101		14.39	11.84	4.98	4.97	0.99	34.65	57.43	6.93	4.95
IT-PaCCSS		Web Data	63,012	Scientific Usage	9.26	8.29	4.62	4.64	1.50	68.8	25.53	4.15	0

Table 1: An overview of the used corpora including domain, corpus size, license, sentence length per complex (comp) and simple (simp) text, word length per complex and simple text, and the proportion of simplification transactions per corpus in percent (S=syntactic, L=lexical, L&S=lexical and syntactical, -L&-S=no lexical nor syntactical, I=identical). A complex-simplified text pair is considered as lexical simplification if new tokens are added to the simplified text or tokens are rewritten in the simplified text. A pair is considered as syntactic simplification if the text is split or joined.

The pre-processing with SpaCy includes sentence-splitting, tokenization, lemmatizing, POS-tagging, dependency parsing, named entity recognition, and generating word embeddings. The SpaCy word embeddings are replaced in this study by pre-trained word embeddings of FastText (Grave et al., 2018) to achieve a higher quality[11]. Unless otherwise stated, this data is used to measure the used features.

3.2.1. Single Features

The single features are grouped into *proportion of part of speech (POS) tags*, *proportion of clauses & phrases*, *length of phrases*, *syntactical*, *lexical*, *word frequency*, *word length*, *sentence length*, and *readability features*. An overview is provided in Table 2.

Proportion of POS Tags Features. Gasperin et al. (2009) and Kauchak et al. (2014) name the proportion of POS tags per sentence as a relevant feature for text simplification. According to Kercher (2013), a higher proportion of verbs in German indicates for instance a simpler text because it might be more colloquial. POS tag counts are normalized by dividing them by the number of tokens per text, as in Kauchak et al. (2014). A list of all used POS tags features is provided in Table 2.

Proportion of Clauses and Phrases Features. Gasperin et al. (2009) and recommend using the proportion of clauses and phrases. The clauses and phrases extend and complex a sentence, so they are often split (Gasperin et al., 2009). The proportion of the clauses and phrases is measured using the dependency tree of the texts and differentiated, as shown in Table 2.

Length of Phrases Features. In a study regarding sentence splitting prediction (Gasperin et al., 2009), the length of noun, verb, and prepositional phrases are used as features because the longer a phrase, the more complex the sentence and the higher the amount of processing.

Syntactic Features. We use six syntactic features, computed based on the SpaCy dependency trees and POS tags. Inspired by Niklaus et al. (2019), we measure whether the head of the text is a verb (Feature 1). If the text contains more than one sentence, at least one root must be a verb. Following Universal Dependencies[12], a verb is most likely to be the head of a sentence in several languages. So, sentences whose heads are not verbs might be ungrammatical or hard to read due to their uncommon structure. Therefore, the feature of whether the head of the sentence is a noun is added (2).

Niklaus et al. (2019) also state that a sentence is more likely to be ungrammatical and, hence, more difficult to read if no child of the root is a subject (3).

According to Collins-Thompson (2014), a sentence with a higher parse tree is more difficult to read, we therefore add the parse tree height as well (4).

Feature (5) indicates whether the parse tree is projective; a parse is non-projective if dependency arcs cross each other or, put differently, if the yield of a subtree is discontinuous in the sentence. In some languages, e.g., German and Czech, non-projective dependency trees are rather frequent, but we hypothesize that they decrease readability.

Gasperin et al. (2009) suggest passive voice (6) as a further feature because text simplification often includes transforming passive to active, as recommended in easy-to-read

[11] This has the disadvantage that the here proposed corpus analysis is only available for languages supported by SpaCy and FastText.

[12] https://universaldependencies.org/docs/en/dep/root.html

text guidelines, because the agent of the sentence might get clearer. Due to different dependency label sets in SpaCy for some languages, this feature is only implemented for German and English.

Lexical Features. Further, six features are grouped into lexical features. The lexical complexity (Feature 1) might be a relevant feature because a word might be more familiar for a reader the more often it occurs in texts. In order to measure the lexical complexity of the input text, the third quartile of the log-ranks of each token in the frequency table is used (Alva-Manchego et al., 2019).

The lexical density –type-token-ratio– (2) is calculated using the ratio of lexical items to the total number of words in the input text (Martin et al., 2018; Collins-Thompson, 2014; Hancke et al., 2012; Scarton et al., 2018). It is assumed that a more complex text has a larger vocabulary than a simplified text (Collins-Thompson, 2014).

Following Collins-Thompson (2014), the proportion of function words is a relevant feature for readability and text simplification. In this study, function words (3) are defined using the universal dependency labels "aux", "cop", "mark" and "case".

Additionally, we added the proportion of multi-word expressions (MWE, 4) using the dependency labels "flat", "fixed", and "compound" because it might be difficult for non-native speakers to identify and understand the separated components of an MWE, especially when considering long dependencies between its components.

The ratio of referential expressions (5) is also added based on POS tags and dependency labels. The more referential expression, the more difficult the text because the reader has to connect previous or following tokens of the same or even another sentence. Lastly, the ratio of named entities (6) is examined because they might be difficult to understand for non-natives or non-experts of the topic.

Word Frequency Features. As another indication for lexical simplification, the word frequency can be used (Martin et al., 2018; Collins-Thompson, 2014). Complex words are often infrequent, so word frequency features may help to identify difficult sentences. The frequency of the words is based on the ranks in the FastText Embeddings (Grave et al., 2018). The average position of all tokens in the frequency table is measured as well as the position of the most infrequent word.

Word and Sentence Length Features. Word length and sentence length are well-established measurements used for readability measurement. Following Scarton et al. (2018), we distinguish word length in number of characters, and syllables and sentence length in number of characters, syllables, and words.

Readability Metric Features. Furthermore, as proposed by Martin et al. (2018), we use readability metrics. Readability metrics calculate based on sentence length and number of syllables the complexity of a text and estimates, for example, the minimum grade of understanding. We differentiate between Flesch-Kincaid Grade Level and Flesh Reading Ease (Kincaid et al., 1975).

3.2.2. Paired Features

The paired features (see Table 3) are grouped into *lexical*, *syntactic*, *simplification*, *word embeddings*, and *machine translation* features.

Lexical Features. Inspired by Martin et al. (2018) and Alva-Manchego et al. (2019), the following proportions relative to the simplified or complex texts are included as lexical features:

- **Added Lemmas:** Additional words can make the simplified sentence more precise and comprehensible by enriching it with, e.g., decorative adjectives or term definitions.
- **Deleted Lemmas:** Deleting complex words might contribute to ease of readability.
- **Kept Lemmas:** Keeping words, on the other hand, might contribute to preserving the meaning of the text (but also its complexity). Kept lemmas describe the words which occur in both texts but might be differently inflected.
- **Kept Words:** Kept Words are a portion of kept lemmas, they describe the proportion of words which occur exactly in the same inflection in both texts.
- **Rewritten Words:** Words which are differently inflected in the simplified text, compared to the complex one, but have the same lemma are called rewritten words. Granted that complex words are rewritten, a higher amount of rewritten words represents a more simplified text.

The compression ratio is similar to the Levenshtein Distance and measures how many characters are left in the simplified text compared to the complex text. The Levenshtein Similarity measures the difference between complex and simplified texts by insertions, substitutions, or deletions of characters in the texts.

Syntactic Features. The idea of the features of split and joined sentences are based on Gasperin et al. (2009), both show an applied simplification transaction. The sentence is counted as split if the number of sentences of the complex text is lower than of the simplified text. The sentence is counted as joined if the number of sentences of the complex text is higher than of the simplified text.

Simplification Features. In order to address more simplification transactions, we measure lexical, syntactical, and no changes. A complex-simplified-pair is considered as a lexical simplification if tokens are added or rewritten in the simplified text. A complex-simplified-pair is considered as a syntactic simplification if the text is split or joined. Also, a change from non-projective to projective, passive to active, and a reduction of the parse tree height are considered as syntactic simplifications. A complex-simplified-pair is considered as identical if both texts are the same, so no simplification has been applied. As each pair is solely analyzed, the standard text simplification evaluation metric SARI (Xu et al., 2016), which needs several gold references, cannot be considered in the analysis.

Word Embedding Features. The similarity between the complex and the simplified text (Martin et al., 2018) is measured using pre-trained FastText embeddings (Grave et al.,

2018). We consider cosine similarity, and also the dot product (Martin et al., 2018). The higher the value, the more similar the sentences, the more the meaning might be preserved and the higher the simplification quality might be.

Machine Translation (MT) Features. Lastly, three MT features are added to the feature set, i.e., BLEU, ROUGE-L, and METEOR. As text simplification is a monolingual machine translation task, evaluation metrics from MT, in particular the BLEU score, are often used in text simplification. Similar to the word embedding features, the higher the value the more meaning of the complex text is preserved in the simplified text. The BLEU score is a well-established measurement for MT based on n-grams. We use 12 different BLEU implementations, 8 from the Python package NLTK and 4 implemented in Sharma et al. (2017).

3.3. Evaluation

The research questions stated in Section 1 will be answered using non-parametric statistical tests using the previously described features on the eight corpora.

In order to answer the first research question regarding differences between the simplified and the complex text, the complexity level is the dependent variable (0: complex, 1: simple). The features previously named are the independent variables and the values per complex-simple pairs are the samples. To evaluate whether the feature values differ between the simplified and complex texts, we use non-parametric statistical hypothesis test for dependent samples, i.e., Wilcoxon signed-rank tests. Afterwards, we measure the effect size r, where r>=0.4 represents a strong effect, 0.25<=r<0.4 a moderate effect and 0.1<=r<0.25 a low effect.

For the analysis of the research questions 2 and 3 regarding differences between the corpora regarding domains or languages, Kruskal–Wallis one-way analyses of variance are conducted. Therefore, the dependent variables are the languages or domains and the independent variables are the paired and difference features. For the analysis within domains and languages, the tests are evaluated against all corpora of one domain or language, e.g., for Wikipedia data the values of EN-QATS and EN-TurkCorpus are analyzed. For the analysis within and across languages and domains, the tests are evaluated against stacked corpora. All corpora assigned to the same language or domain are stacked to one large corpus, e.g., the German corpus and IT-PaCCSS are stacked as web data corpus and are tested against the stacked Wikipedia corpus and the stacked news article corpus. If there is a significant difference between the groups, a Dunn-Bonferonni Post-hoc Test is applied to find the pair(s) of the difference. Afterwards, again, the effect size is measured using the same interpretation levels as for the Wilcoxon signed-rank tests.

4. Results and Discussion

The results of the analysis are reported on eight corpora, five languages, three domains, and 104 features using Wilcoxon signed-rank tests and Kruskal-Wallis tests[13]. These results should be handled with caution because they might be biased due to errors in SpaCy's output, e.g., regarding dependency parsing and named entity recognition, or due to the unbalanced corpora.

4.1. Differences between Complex and Simplified Texts (RQ1)

The results concerning the question whether the feature values of complex texts and its simplified version differ significantly are summarized in Table 2.

For all three sentence length features, both readability features, and the parse tree height feature, Wilcoxon signed-rank tests indicate at least low but significant effects between the complex and simplified text pairs overall all corpora when analyzing the corpora solely.

The result is not surprising since sentence length has already been shown to be a relevant feature in different languages, e.g., in English Napoles and Dredze (2010) and Martin et al. (2018), in German Hancke et al. (2012), and in Portuguese Aluisio et al. (2010).

The parse tree height also differs significantly for all corpora in the complex and simplified texts. Pilán and Volodina (2018) and Napoles and Dredze (2010) also conclude in their studies regarding Swedish and English that the parse tree height is a relevant complexity measurement feature.

Considering differences between the proportion of verbs in complex and simplified texts, the Wilcoxon signed-rank tests indicate at least low but significant effects for each corpus except EN-QATS. So, the assumption of Kercher (2013), that a higher number of verbs simplifies a text can be generalized to other languages than German.

In contrast, several features are only relevant for a few corpora and differ even more in the effect size. For example, Wilcoxon signed-rank tests indicate a strong significant effect for the lexical density in EN-Newsela-2015 (M_{comp}=0.89±0.08, M_{simp}=0.93±0.07, n=141,582, t(141,581)=1329762920.5, p <=.01) but only indicate at most moderate effects on three other corpora and no effect on the remaining four corpora. Furthermore, for several features, the Wilcoxon signed-rank tests indicate no significant difference not even for one corpus, e.g., non-projectivity, proportion of symbols, or proportion of named entities (see Table 2).

Overall, the results show that some of the proposed features help to explain the simplification processes in the selected corpora even if the features might well not be sufficient to explain the simplification process at all. In the next Subsections, we will follow up on these assumptions by comparing the consistency of the simplification process regarding domains and languages.

4.2. Domain Simplification Consistency (RQ2)

Since the selected features are useful to explain the simplification process, the consistency or differences in the simplification process are measured using the difference version of these features as well as the paired features. The results regarding domains are separated into differences within and across domains.

[13] All statistical characteristics are provided as supplementary material in the linked github repository.

Prop. POS Tags			Prop. Clauses & Phrases			Readability			Lexical Features		
Adjectives	♡■		All Clauses	♡♠■		FRE	♥♠		Lex. Complexity	♡♠■■	■
Adpositions	♡♠■		Coord. clauses	♡		FKGL	♥♠■		Lex. Density	♡	♣
Adverbs	♡♠■		Subord. clauses	♡		**Word Length in**			Prop. MWEs	♠■	
Auxiliary verbs	♡♠■		PPs	♡♠■		Characters	♥♠■	♣	Prop. Named Ent.	♠	
Conjunctions	♡♠■		Relative Phrases	♡♠■		Syllables	♥♠		Prop. Funct. Words	♡♠■	
Determiners	♡♠■		**Syntactic Features**			**Length of Phrases**			Prop. Ref. Expr.		
Interjections			Head is Noun			Noun Phrase	♡		**Word Frequency**		
Nouns	♡		Head is Verb			Verb Phrase		♣	Avg. Position	♡♠■	♣
Numerals	♡♠■		Subj. child of root			PPs			Max. Position	♡♠■	
Particles	♡		Parse Tree Height	♥♠■	♣	**Sent. Length in**					
Pronouns	♡♠■	♣	Non-Projectivity		♣	Characters	♥♠■	♣			
Punctuation	♡♠■	♣	Passive Voice			Syllables	♥♠■				
Symbols						Words	♥♠■				
Verbs	♥♠■										

Table 2: The single and difference features are presented sorted by groups. In the 2nd, 5th, 8th and 11th column, differences between the complex-simplified pairs are listed: ♥ symbolizes differences in the pairs per corpus (RQ1), ♣ in the pairs within domains, ♠ in the pairs across domains, ♦ in the pairs within languages, and ■ in the pairs across languages. In the 3rd, 6th, 9th and 12th column, the differences between the languages and the domains are shown in across and within settings using the same symbols. The color of the symbols indicates the distribution of the effects: Black illustrates an effect for all languages or domains, gray for most of them and lightgray/white for only a few.

Lexical	Effect	Simplification	Effect
Prop. Added Lemmas	♣	Lexical Simplification	
Prop. Deleted Lemmas		Syntactic Simplification	
Prop. Kept Lemmas	♣	Identical	
Prop. Kept Words	♣	**Machine Translation**	**Effect**
Prop. Rewritten Words		BLEU	♠ ♣
Compression Ratio		METEOR	
Levenshtein Similarity		ROUGE-L	♣
Levenshtein Distance		**Word Embeddings**	**Effect**
Syntactic	**Effect**	Cosine Similarity	
Sentence Split		Dot Product	
Sentences Joined			

Table 3: The paired features are presented sorted by their group label. The significant effects per features are highlighted using the following symbols per research question: The ♣ symbol represents within domain results, ♠ across domains, ♦ within languages, and ■ across languages. Black illustrates an effect for all languages or domains, gray for most of them and white for only a few.

Within Domains. When the features are analyzed regarding the consistency within a domain, significant differences are indicated only between the corpora of the web text domain. The German and Italian corpora of this domain differ significantly with a low effect for 14 features (see Table 2 and Table 3), e.g., parse tree height difference, difference of non-projectivity, characters per word, and BLEU score.
The parse tree height is significantly more reduced in German (Difference: M_{DE}=1.16±1.96, N_{DE}=1,888) than in Italian (M_{IT}=0.13±0.64, N_{IT}=63,012, H(1)=759.71, p <=.01, r=.11) which might be due to a higher average parse tree height in the German corpus (M_{comp}=4.74±2.41, M_{simple}=3.58±1.35) than in the Italian corpus (M_{comp}=3.14±0.97, M_{simp}=3.02±0.94).
Parse tree height and sentence length are reduced in both corpora in the simplified texts, but, surprisingly, the average word length in characters is slightly increased in Italian (M_{comp}=4.62±0.91, M_{simp}=4.64±0.89). So, this effect might explain the significant difference between both corpora and should be considered for following analysis.
Overall, the differences between the two web data corpora may tie to the high proportion of only lexical simplification in the IT corpus and high proportion of lexical and syntactic simplification in the DE corpus. The other corpora within one domain are more similar in their distribution, which may explain why they do not differ significantly.

Across Domains. The only significant difference across all domains is the BLEU score (H(2)=1429.0979, p <=.01, r=.12). A Dunn-Bonferonni Post-hoc Test indicates that the web (M=0.61±0.15, N=64,900) and Wikipedia data (M=0.67±0.22, N=19,377) are differing. This confirms the findings of Sulem et al. (2018) that BLEU is not suitable for measuring text simplification.
Furthermore, the domains differ also in more features even if not significantly between all domains. The following features show only a significant difference between complex and simplified texts in one of the domains.

- **web data:**
  - word frequency avg. position (r=.26, p <=.01),
  - word frequency max. position (r=.14, p <=.01),
  - prop. of adjectives (r=.22, p <=.01),
  - prop. of adverbs (r=.21, p <=.01),
  - prop. of determiners (r=.52, p <=.01),
  - prop. of function words (r=.31, p <=.01), and
  - prop. of numerals (r=.18, p <=.01)
- **newspaper articles:**
  - prop. of clauses (r=.15, p <=.01),
  - prop. of MWEs (r=.14, p <=.01),
  - prop. of adpositions (r=.12, p <=.01),
  - prop. of conjunctions (r=.2, p <=.01),
  - prop. of propositional phrases (r=.12, p <=.01),
  - prop. of relative phrases (r=.16, p <=.01).

In contrast, some features are relevant for text simplification in all domains, i.e., characters per sentence, syllables per sentence, words per sentence, parse tree height, proportions of auxiliary verbs and of verbs, FKGL, and FRE.
Overall, these results show, also in combination with BLEU as the only significant difference across domains, that the simplification process seems to be consistent across the web, Wikipedia, and news article domain.

4.3. Language Simplification Consistency (RQ3)

The results of the differences in the simplification process regarding languages are separated into differences within and across languages.

Within Languages. The comparison within a single language is done only for English because this is the only language where we have more than one corpus. All English corpora[14] are combined into a large corpus of 230,144 complex-simplified pairs. Using a Kruskal–Wallis test, no significant difference is indicated between the English corpora, which led to the conclusion that the simplification process measured using several linguistic features in these corpora is consistent. However, this must be handled with particular caution because the size of the corpora is unbalanced and, furthermore, the simplification processes applied have different focuses, varying between lexical and syntactic simplification, e.g., EN-QATS has 15.45% of syntactically simplified text pairs whereas EN-Newsela-15 has only 0.03% (see Table 1).

Across Languages. The only significant difference between all languages is the lexical complexity difference ($H(4)$=425.1521, p <=.01, r=.12). A Dunn-Bonferonni Post-hoc Test indicates that only the German (M=0.33±1.07) and the Czech corpus (M=-0.09±1.19) are significantly differing. Surprisingly, the lexical complexity seems to increase in Czech during simplification.
On the one hand, Wilcoxon signed-rank tests also indicate some features with a significant difference in the language-wise data regarding complex and simplified texts for only one or two languages:

- **DE:** lexical complexity (r=.31, p <=.01),
- **IT:** proportion of function words (r=.32, p <=.01), proportion of numerals (r=.19, p <=.01),
- **DE and IT:** proportion of pronouns (r_{DE}=.31, r_{IT}=.31, p <=.01),
- **EN and CS:** proportion of relative phrases (r_{CS}=.12, r_{EN}=.15, p <=.01).

On the other hand, the simplification processes of all languages are similar regarding the following 9 features: characters per sentence, syllables per sentence, words per sentence, parse tree height, proportion of adpositions, proportion of verbs, proportion of prepositional phrases, FKGL, and FRE. Following these results as well as the result of the lexical complexity as sole difference regarding languages, the simplification process seems to be more or less consistent across Czech, German, English, Spanish, and Italian.

[14] From EN-Newsela-2016 only level 0 to 1 is used.

5. Conclusion and Future Works

This study investigated whether text simplification processes differ within or across five languages (Czech, German, English, Italian, and Spanish) and three domains (newspaper articles, web texts, and Wikipedia texts). To this end, we first tested linguistic features as to their relevance for characterizing the differences in complex-simplified text pairs of eight corpora. Statistical tests indicate significant differences for some of the features, e.g., sentence length, parse tree height, or proportion of verbs. So, these features are used to measure the simplification process in this study. However, the selected features might well not be sufficient to explain the whole simplification process. Other features, such as morphological or grammatical features could improve it in future work.
Furthermore, our study shows differences in the relevance of features per corpus. This insight was further refined regarding differences within and across domains. For the newspaper and Wikipedia corpora, no differences were found within each of the two domains, the statistical tests indicated only differences for the web corpora. These results as well as the finding of only one differing feature across domains, led to the assumption that the simplification process is consistent across and within domains, such as similarly stated in Vajjala and Meurers (2014).
Our study regarding within and across language comparisons also supports the results of Scarton et al. (2017) and Finnimore et al. (2019): text simplification seems to be consistent across languages, which indicates that cross-lingual text simplification based on a single language-independent feature set is a viable approach. Nevertheless, features might be weighted differently per language.
Overall, the negative statistical tests regarding differences across and within domains and languages led to the assumption that the simplification process is robust across and within domains and languages. Especially the features of parse tree height, readability, and sentence length seem to be robust against domains and languages. In contrast, in the evaluation and designing of text simplification models, features such as lexical complexity, and BLEU score should be used with caution due to their found differences in the corpora. These findings might help to build a text simplification model or a text simplification metric that is aware of language or domain characteristics.

6. Acknowledgments

This research is part of the PhD-program "Online Participation", supported by the North Rhine-Westphalian funding scheme "Forschungskolleg".

7. Bibliographical References

Aluisio, S., Specia, L., Gasperin, C., and Scarton, C. (2010). Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 5th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Los Angeles, California, June. ACL.

Alva-Manchego, F., Martin, L., Scarton, C., and Specia, L. (2019). EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference*

on EMNLP and the 9th IJCNLP, pages 49–54, Hong Kong, China, November. ACL.

Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *ITL - International Journal of Applied Linguistics*, 165(2):97–135.

Finnimore, P., Fritzsch, E., King, D., Sneyd, A., Ur Rehman, A., Alva-Manchego, F., and Vlachos, A. (2019). Strong baselines for complex word identification across multiple languages. In *Proceedings of the NAACL HTL 2019*, pages 970–977, Minneapolis, Minnesota, June. ACL.

Gasperin, C., Specia, L., Pereira, T. F., and Aluisio, R. M. (2009). Learning when to simplify sentences for natural text simplification. In *Proceedings of ENIA*, pages 809–818.

Hancke, J., Vajjala, S., and Meurers, D. (2012). Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012*, pages 1063–1080, Mumbai, India, December. The COLING 2012 Organizing Committee.

Kauchak, D., Mouradi, O., Pentoney, C., and Leroy, G. (2014). Text simplification tools: Using machine learning to discover features that identify difficult text. In *2014 47th Hawaii International Conference on System Sciences*, pages 2616–2625, Jan.

Kercher, J. (2013). *Verstehen und Verständlichkeit von Politikersprache*. Springer Fachmedien Wiesbaden.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Martin, L., Humeau, S., Mazaré, P.-E., de La Clergerie, É., Bordes, A., and Sagot, B. (2018). Reference-less quality estimation of text simplification systems. In *Proceedings of the 1st Workshop on Automatic Text Adaptation*, pages 29–38, Tilburg, the Netherlands, November. ACL.

Napoles, C. and Dredze, M. (2010). Learning simple Wikipedia: A cogitation in ascertaining abecedarian language. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing*, pages 42–50, Los Angeles, CA, USA, June. ACL.

Niklaus, C., Freitas, A., and Handschuh, S. (2019). MinWikiSplit: A sentence splitting corpus with minimal propositions. In *Proceedings of the 12th INLG*, pages 118–123, Tokyo, Japan, October–November. ACL.

Paetzold, G., Alva-Manchego, F., and Specia, L. (2017). MASSAlign: Alignment and annotation of comparable documents. In *Proceedings of the IJCNLP 2017*, pages 1–4, Tapei, Taiwan, November. ACL.

Pilán, I. and Volodina, E. (2018). Investigating the importance of linguistic complexity features across different datasets related to language learning. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 49–58, Santa Fe, New-Mexico, August. ACL.

Scarton, C., Palmero Aprosio, A., Tonelli, S., Martín Wanton, T., and Specia, L. (2017). MUSST: A multilingual syntactic simplification tool. In *Proceedings of the IJCNLP 2017*, pages 25–28, Tapei, Taiwan, November. ACL.

Scarton, C., Paetzold, G., and Specia, L. (2018). Text simplification from professionally produced corpora. In *Proceedings of the 11th LREC*, Miyazaki, Japan, May. ELRA.

Sharma, S., El Asri, L., Schulz, H., and Zumer, J. (2017). Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *CoRR*, abs/1706.09799.

Sulem, E., Abend, O., and Rappoport, A. (2018). BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on EMNLP*, pages 738–744, Brussels, Belgium, October-November. ACL.

Vajjala, S. and Meurers, D. (2014). Readability assessment for text simplification:from analyzing documents to identifying sentential simplifications. *International Journal of Applied Linguistics, Special Issue on Current Research in Readability and Text Simplification*.

Yimam, S. M., Štajner, S., Riedl, M., and Biemann, C. (2017). Multilingual and cross-lingual complex word identification. In *Proceedings of RANLP 2017*, pages 813–822, Varna, Bulgaria, September. INCOMA Ltd.

8. Language Resource References

Barančíková, P. and Bojar, O. (2019). COSTRA 1.0: A dataset of complex sentence transformations. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Brunato, D., Cimino, A., Dell'Orletta, F., and Venturi, G. (2016). PaCCSS-IT: A parallel corpus of complex-simple sentences for automatic text simplification. In *Proceedings of the 2016 Conference on EMNLP*, pages 351–361, Austin, Texas, November. ACL.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the 11th LREC*, Miyazaki, Japan, May. ELRA.

Klaper, D., Ebling, S., and Volk, M. (2013). Building a German/simple German parallel corpus for automatic text simplification. In *Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19, Sofia, Bulgaria, August. ACL.

Štajner, S., Popović, M., Saggion, H., Specia, L., and Fishel, M. (2016). Shared task on quality assessment for text simplification. In *qats2016: LREC 2016 Workshop & Shared Task on Quality Assessment for Text Simplification (QATS), 28th May 2016, Portorož, Slovenia ; proceedings*, pages 22–31, Paris. ELRA-ERDA. Online-Ressource.

Xu, W., Callison-Burch, C., and Napoles, C. (2015). Problems in current text simplification research: New data can help. *Transactions of the ACL*, 3:283–297.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., and Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the ACL*, 4:401–415.