How to Make Troubleshooting Simpler? Assessing Differences in Perceived Sentence Simplicity by Native and Non-native Speakers

Sanja Štajner
Data and Web Science Group
University of Mannheim, Germany
sanja@informatik.uni-mannheim.de

Abstract

Text Simplification (TS) aims to make texts more accessible to various readers by reducing reading time and/or by facilitating understanding and access to the relevant information. In this study, we address two topics which are important for TS but have not received much attention so far. First, we explore which types of syntactic simplification transformations make troubleshooting in IT domain easier for both native and non-native English speakers. Second, we explore how the choice of TS evaluation strategy influences final results. Our experiments show that grammaticality of a sentence influences the perceived simplicity by native and by non-native speakers differently. We also find that a high inter-annotator agreement can be achieved only in the case of the relative assessment of the sentence pairs in which one sentence is significantly simpler than the other one.

Keywords: evaluation, text simplification, sentence simplicity

1. Introduction

Text simplification (TS) has the goal of making texts more accessible by reducing reading time and/or improving understanding of the information contained in them. So far, TS systems have been proposed for many languages, e.g. English (Carroll et al., 1999; Coster and Kauchak, 2011a; Siddharthan and Angrosh, 2014; Štajner and Glavaš, 2017; Nisioi et al., 2017), Spanish (Saggion et al., 2015; Štajner et al., 2015b), Portuguese (Aluísio and Gasperin, 2010; Specia, 2010), Italian (Barlacchi and Tonelli, 2013), Basque (Aranzabe et al., 2012). The proposed TS systems had various target populations in mind, e.g. children (Barlacchi and Tonelli, 2013), people with low literacy levels (Aluísio and Gasperin, 2010), non-native speakers (Paetzold and Specia, 2016b), and people with various cognitive or reading impairments (Saggion et al., 2015; Rello et al., 2013; Orasan et al., 2013). The majority of the proposed TS systems focused on simplifying either news articles, or Wikipedia articles, or both. The latest state-of-the-art TS systems (Nisioi et al., 2017; Zhang and Lapata, 2017) use neural architectures and are trained on the English Wikipedia – Simple English Wikipedia (EW–SEW) TS datasets (Coster and Kauchak, 2011b; Hwang et al., 2015). By being fully supervised and trained on the EW–SEW TS dataset, they represent the 'general' TS systems which should make texts more accessible to everyone.

1.1. Evaluation of TS systems

In spite of the recent increased interest in text simplification, there are no common standards in evaluation of TS systems. Ideally, TS systems should be evaluated at the text level, by measuring reading time and understanding by the final users. However, such an evaluation is time-consuming, and in the case of vulnerable target populations, the access to the final users might be difficult. Therefore, in practice, TS systems are usually evaluated at the sentence level by native or non-native speakers, or a mixture of both. This already raises some important issues, as randomly-selected or crowdsourced evaluators might not be good proxies for the target populations. Apart from that, such human evaluations have varied from one study to another with regards to the number of annotators, the type of annotators (native vs. non-native), the type of evaluation (absolute score vs. relative comparison), the evaluation scale (0/1, 1–3, or 1–5), etc. (see Table 1 in Section 2), which brings additional problems in comparing the performances of TS systems from different studies.

1.2. Goals and Contributions

This work has two main goals. The first is to better understand how different syntactic transformations and the grammaticality of a sentence influence the perceived sentence simplicity by native and non-native English speakers. The second goal is to explore how the type of annotators and the type of evaluation influence final results and the inter-annotator agreement (IAA). To achieve those goals, we explore the following research questions:

- **RQ1**: How does the grammaticality of a text snippet influence its simplicity and whether this influence vary depending on the type of evaluators (native vs. non-native speakers) or not?

- **RQ2**: Is the absolute simplicity of a text snippet perceived differently by native and by non-native speakers?

- **RQ3**: Is there a difference in simplification gain after applying particular simplification operations on a text snippet (i.e. its relative simplicity) depending on whether it is evaluated by a native or by a non-native speakers?

- **RQ4**: How does the type of evaluators (native vs. non-native speakers) and the evaluation type (absolute vs. relative simplicity) reflect on the inter-annotator agreement?

Study – Language	Simp.type		Readab.	MTeval	Cover.	Human evaluation of sentence/word simplicity							
	Synt.	Lex.				Native	1–5	1–3	0–1	Rel.	#annot.	#sent.	mod.
(Specia, 2010) – PT	+	+	–	+	–	?	–	+	–	–	?	20	+
(Yatskar et al., 2010) – EN	–	+	–	–	+	+	–	+	–	–	3	100*	–
(Coster and Kauchak, 2011a) – EN	+	+	–	+	–	–	–	–	–	–	–	–	–
(Wubben et al., 2012) – EN	+	+	+	+	–	–	+	–	–	–	46	20	+
(Glavaš and Štajner, 2013) – EN	+	–	+	–	–	–	–	+	–	–	3	70	?
(Angrosh et al., 2014) – EN	+	+	–	–	–	±?	+	–	–	–	?	50	+
(Saggion et al., 2015) – ES	+	+	+	–	–	+	+	+	–	–	25	48	+
(Baeza-Yates et al., 2015) – ES	–	+	–	–	+	+	–	–	+	–	3	200*	+
(Štajner et al., 2015b) – ES	+	+	–	+	–	±	+	–	–	+	13	20-40	+
(Glavaš and Štajner, 2015) – EN	–	+	–	–	+	–	+	–	–	+	2	80	–
(Paetzold and Specia, 2016b) – EN	–	+	–	–	+	–	–	–	–	–	–	–	–
(Xu et al., 2016) – EN	+	+	–	+	–	?	–	–	–	+	5	?	?

Table 1: The types of evaluations used in various text simplification studies ('*' signifies the number of examples/words instead of the number of sentences; '?' that the answer cannot be found in the paper; and '±?' that the evaluators are most probably a mixture of native and non-native speakers).

To better place our work into the current TS literature and clarify our research goals and contributions, we present the most relevant related work in Section 2. In Section 3, we describe the procedure for collecting the new dataset used in this study. The next four sections (Sections 4–7) describe the experimental setup and provide results and detailed discussions for each of the four research questions (RQ1–4) separately. Section 8 summarises the most important results and discuss their potential impact on TS research.

2. Related Work

In this section, we present the most relevant related work with regards to the automatic evaluation methods in TS (Section 2.1), human evaluation procedures in TS (Section 2.2), inter-annotator agreement (Section 2.3), and the existing evaluation datasets (Section 2.4).

2.1. Automatic Evaluation in TS

The main way of evaluating text simplification (TS) systems, either manual or automated, is by human evaluation. Some studies, however, additionally include an automatic evaluation of TS systems, either by using readability formulae (e.g. (Saggion et al., 2015; Drndarević et al., 2013; Woodsend and Lapata, 2011)), or in the case of machine translation (MT) based TS (e.g. (Specia, 2010; Xu et al., 2016)), by using some of the automatic machine translation evaluation measures such as BLEU (Papineni et al., 2002) or NIST score (Doddington, 2002). Although such automatic evaluation measures allow for a greater number of instances to be evaluated than by means of human evaluation, they still have a number of shortcomings. Readability metrics are reliable only at the text level and not at the sentence level (and are oblivalent to grammaticality and meaning), while machine translation evaluation metrics, although showing good correlation with some human judgements (Wubben et al., 2012; Štajner et al., 2014; Xu et al., 2016; Štajner et al., 2016), do not reward for TS-specific transformations such as sentence shortening or strong paraphrasing, and do not take into account the input sentences (Štajner et al., 2015a; Xu et al., 2016; Štajner and Nisioi, 2018). Additionally, the MT-evaluation metrics require 'gold standard' simplifications which are rarely available in TS.

The systems which only perform lexical simplification (no syntactic simplification) are usually evaluated automatically by the number of changes performed and the coverage of changes over a 'gold standard' dataset of complex words for non-native speakers (Horn et al., 2014; Glavaš and Štajner, 2015; Paetzold and Specia, 2016a; Paetzold and Specia, 2016b). This type of evaluation is thus limited to those particular test instances.

2.2. Human Evaluation in TS

The main and most reliable method for evaluating TS systems is still considered to be human evaluation. Although it should ideally be performed at the text level and by the final target users, given the time such evaluation would require and difficulties to reach some of the target populations, human evaluation of TS systems is usually done only at the sentence level, by either native or non-native speakers (or, less often, a mixture of both). The annotators are asked to rate the simplified sentences for their grammaticality, meaning preservation, and simplicity, either on an absolute scale or in a pairwise comparison. The problem is that different TS studies use different evaluation strategies and thus it is not possible to compare their results. Table 1 presents an overview of different types of evaluation strategies used for assessing the simplicity of TS output so far. As can be seen, human evaluation is sometimes performed in terms of an absolute simplicity score on a 1–5 or 1–3 level scale, sometimes as a 0–1 labeling (simpler/not simpler than the original), and sometimes as a relative pairwise comparison of the output of different systems or system configurations (*Rel.*). The number of annotators (*#annot.*), and the number of sentences (*#sent.*) also vary, as well as whether the evaluation is performed only on the sentences which underwent some change (*mod.*) or a random subset of sentences (including the unchanged ones).

2.3. Inter-Annotator Agreement

Although Yatskar et al. (2010) reported a better inter-annotator agreement among native than among non-native evaluators on a 1–5 level evaluation task, no other stud-

Answer type	Text
Original (O)	Your payment received not and application submitted
Grammatical (G)	Your payment **was not received** and **the** application **was** submitted.
Split (S)	Your payment was not received**. However,** the application was submitted.
All (A)	**We did not receive** your payment. **But, we received your** application.

Table 2: An example of the original answer (O), grammatically corrected answer (G), answer with split sentences (S), and the answer with applied *additional operations* (A). Differences between the answers are shown in bold. The corresponding question was: *Was my payment received on time?*

ies tried to further explore the influence of the annotators type on the obtained evaluation results. The differences in the obtained results could be particularly pronounced in the case of relative comparisons, as the original sentence might already be simple enough for native speakers and therefore, its further simplification would not be rewarded by native annotators as much as by non-native annotators. Another problem could be that the original sentence might be so complex for the non-native annotators that any of the performed simplification transformations (usually very few in automatic text simplification systems) could not make it any simpler.

Furthermore, it has never been explored how much the grammaticality of a sentence influences its perceived simplicity. This might be an important factor, as the current state-of-the-art TS systems still produce many ungrammatical sentences. An output sentence which is rated as complex might be rated that way for various reasons: (1) because it is ungrammatical (in spite of correctly applied lexical or syntactic simplification operations); (2) because it was left unchanged and the original sentence was already complex; (3) because the simplification operations were incorrectly applied and led to generating a more complex sentence (which might be completely grammatical). Those three cases could be differently rated by native and by non-native annotators, as for example, the native evaluators might penalize the ungrammaticality more severely, while the non-native annotators might reward lexical and syntactic simplicity regardless of the grammaticality.

2.4. Existing Datasets for TS Evaluation

The existing datasets with human evaluation scores (e.g. those systems shown in Table 1) are not convenient for assessing the influence of grammaticality and various sentence simplification operations on the perceived simplicity of a text snippet, as they do not control for one variable at the time. The majority of those datasets only contain those sentences which underwent at least one syntactic or lexical transformation, and such sentences are often ungrammatical. By using such a dataset we would not be able to know whether the change in simplicity score between the original and automatically simplified (ungrammatical) sentence comes from the fact that the grammaticality of the sentence was damaged, or from the changes that were made (and which kind of changes were made exactly), while all those factors may play a role. Additionally, only one of those datasets (Štajner et al., 2015b) contains the evaluation scores assigned by both native and non-native speakers and would thus allow for comparison of scores obtained by native vs. non-native evaluators. Yet again, that dataset does not have a sufficient number of sentences to allow controlling for the grammaticality and the sentence transformation type. To control for all those factors at the same time, we create a new dataset described in the next section.

3. Data Collection

We collect questions and answers (Q&A) from an IT service in a hospital in India and from the WMT 2016 shared task for the IT domain.[1] We opt for having a Q&A type of dataset to have a larger context for better simplicity assessment, and for testing the previously mentioned issues in a real-world scenario, in which a reader is required to understand the answer and find the necessary information. Out of all Q&A from those two sources, we selected 30 (20 from the first source and 10 from the second) which fulfilled the following two conditions: (1) the answer was grammatically incorrect (containing more than one grammatical error) and (2) the answer was sufficiently complex to allow applying sentence splitting and at least one of the following simplification operations (*additional transformations*):

- removing superfluous words
- conversion of passive to active voice
- disambiguation of meaning
- conversion to the canonical subject-verb-object form

As *additional transformations* we choose the most frequently used operations in various guidelines for producing easy-to-read texts (PlainLanguage, 2011; Mencap, 2002; Freyhoff et al., 1998) and in rule-based text simplification modules (Siddharthan and Angrosh, 2014; Saggion et al., 2015). We do not focus on lexical simplification on purpose, as LS systems are usually evaluated for their coverage on the existing, specially designed datasets (Horn et al., 2014; Glavaš and Štajner, 2015; Paetzold and Specia, 2016a; Paetzold and Specia, 2016b). Instead, we focus on the simplification operations which operate on a syntactic or discourse level.

For each of the 30 answers, in addition to the original answer (O), we produce three simplified versions: (1) the grammatically corrected version (G); (2) the sentence-split version (S) by performing sentence splitting on the grammatically corrected version as many times as possible to satisfy the main rule of easy-to-read texts (keeping the sentences as short as possible and covering only one main idea per sentence); and (3) the (grammatically corrected) sentence simplified by using sentence splitting and as many as

[1] http://www.statmt.org/wmt16/it-translation-task.html

possible *additional transformations* (A). All three versions were produced by a linguist, native English speaker, and checked by another. Both editors were experienced in manual text simplification. An example of the original sentence (O), its manually corrected version (G), and the two manually simplified versions: after the sentence splitting (S) and after performing additional simplification transformations (A) is presented in Table 2.

The 30 Q&A were divided into two sets of 15 Q&A randomly. Next, we ask 40 native and 40 non-native English speakers to rate each of the four answers for the given 15 questions on a 1–5 Likert scale by saying how easy to understand was the answer (1 → very simple; 5 → very complex). This way, each answer is evaluated by 20 native and 20 non-native English speakers, and each evaluator only evaluates 15 answers. In this way, we follow the common TS evaluation practices of not giving more than 20 tasks/items to the evaluators in the case of crowdsourced evaluation, to avoid the fatigue effect.

Different versions of the same answers were always shown one after another in random order, always together with their corresponding question. We opted for this way of presenting answers, which allows for their direct comparison and ranking, but we did not instruct the annotators to rank them. Instead, we asked them to evaluate them independently of each other. We chose this evaluation setup as it is the most common way of evaluating TS systems.

The native English speakers were from the UK, USA, and Australia. The non-native English speakers were from India and Germany. The English proficiency of non-native speakers was not checked by any kind of qualification tests. However, the language used at their workplace is English. All 80 evaluators (40 native and 40 non-native English speakers) are familiar with the IT support procedures, as they all use computers in their daily tasks at work and have contacted the IT support at least once before. We did not collect any additional information about the evaluators (e.g. gender, age, or name).

4. Influence of Grammaticality (RQ1)

To explore how the grammaticality of an answer influences its perceived simplicity, we conduct two analyses.

First, we calculate the difference between the simplicity score assigned to the grammatically corrected answer and the one assigned to the original answer (where positive number indicates higher simplicity of the grammatically corrected version) for each answer and for each annotator (a total of 30 × 20 = 600 data points). The first row in Table 3 presents the average difference with standard deviation. We find that grammaticality has a significantly (Mann-Whitney U test in SPSS; $p < 0.05$) stronger influence on perceived simplicity within the native speakers than within the non-native speakers.

Second, we calculate the percentage of cases in which the grammatically correct version (G) was rated as simpler than the original (ungrammatical) version (O), and the percentage of cases in which both versions (O and G) were rated as equally simple (the last two rows in Table 3).

The results indicate that the native speakers penalize the simplicity score of ungrammatical sentences significantly more than the non-native speakers do.

Measure	Native	Non-native
Average difference ± st.dev.	0.50 ± 1.07	0.39 ± 1.20
G simpler than O	44.50%	37.50%
G equally simple as O	43.60%	45.83%

Table 3: The average difference (with standard deviation) between the simplicity score of the grammatically correct answer (G) and the simplicity score of the original answer (O), where positive values signify that the original answer was perceived as more complex, and the percentage of cases (out of 600) in which G was rated as simpler than O, or as equally simple as O.

Type	Native	Non-native
Original	2.65 ± 1.12	2.66 ± 1.22
Grammatically correct	2.14 ± 0.99	2.27 ± 1.09
With splitting	**1.95 ± 0.91**	**2.15 ± 1.10**
With addition. transform.	**1.82 ± 0.91**	**2.14 ± 1.20**

Table 4: The mean value of the simplicity score (with standard deviation) for different variants of the answers (the lower the score, the simpler the answer). Statistically significant differences (Mann-Whitney U test in SPSS; $p < 0.02$ and $p < 0.001$, respectively) between the two groups of evaluators (native and non-native) are presented in bold.

5. Absolute Simplicity (RQ2)

The influence of the annotators group (native vs. non-native) on the perceived absolute simplicity of the answer (before and after various modifications) is presented in Table 4. Although both groups start from the similar average simplicity scores for the original answers (2.65 and 2.66, respectively), after the application of simplification transformations (both sentence splitting and additional transformations), the native speakers perceive the simplified answers significantly simpler than the non-native speakers do. Additionally, we notice a lower standard deviation in the scores within the native speakers. This indicates a greater homogeneity within the native than within the non-native annotators, which is in line with the findings of Yatskar et al. (2010) and Yimam et al. (2017) on similar tasks.

6. Relative Simplicity (RQ3)

To investigate the influence of the annotators group (native vs. non-native) on the results of the relative simplicity assessment, for each of the 600 data points (30 Q&A × 20 annotators), we calculate the difference between the scores obtained for two different versions of the same answer. The mean values and standard deviations are presented in Table 5, where positive score indicates that the second version is simpler. To investigate whether there are significant differences in simplicity scores assigned to the answer before and after certain simplification/correction operation, within the same group of annotators, we compare the scores using marginal homogeneity test for two related samples in SPSS, following the methodology for the relative simplicity assessment used by Štajner et al. (2015b).

Comparing	Native	Non-native
Original−Grammatical	***0.50 ± 1.07**	***0.39 ± 1.20**
Grammatical−Split	*0.19 ± 1.01	*0.12 ± 1.07
Grammatical−Additional	***0.32 ± 1.25**	***0.13 ± 1.17**
Original−Split	***0.70 ± 1.20**	***0.51 ± 1.31**
Original−Additional	***0.82 ± 1.32**	***0.52 ± 1.37**
Split−Additional	*0.13 ± 1.12	0.01 ± 1.10

Table 5: The average change (with standard deviation) in simplicity score between the two versions of the same answer. Statistically significant (Mann-Whitney U test in SPSS; $p < 0.05$) differences between the two groups of evaluators are presented in bold. Statistically significant (marginal homogeneity test for two repeated samples in SPSS; $p < 0.01$) differences in simplicity scores between the two versions of the same answer, within the same group of annotators, are marked with an '*'.

For almost all relative comparisons, we find a significant difference in the obtained simplicity gain between the two groups of annotators. The differences are mostly influenced by the way the grammaticality and the application of *additional transformations* change the score. Sentence splitting seems to influence simplicity gain similarly in both groups of annotators. The difference in simplicity gain after applying sentence splitting and multiple transformations on the grammatically corrected versions (*Grammatical−Additional*) is almost three times more pronounced within the native than the non-native annotators.

The results also indicate that the grammaticality of a sentence has much stronger influence on the simplicity gain than any simplification transformation. This calls for attention in current practices in text simplification evaluation. As we saw earlier (Table 1), some studies evaluate only the sentences which have been changed, while the others also evaluate the unchanged sentences. Transformed sentences are often ungrammatical, and according to our results, this can influence the perceived simplicity more than any simplification operation, thus blurring the TS evaluation results. Furthermore, the application of multiple transformations on the already short sentences (*Split−Additional*) seems not to have much influence, on average, in any of the two annotator groups.

The large standard deviations within the groups indicate a high heterogeneity in perceived simplicity gain within both groups, indicating the need for having a large number of annotators for a reliable evaluation of simplicity gain (relative comparisons).

7. Inter-Annotator Agreement (RQ4)

κ range	Original	Gramm.	Split	Addit.
$0 < \kappa \leq 0.2$	139	168	158	152
$0.2 < \kappa \leq 0.4$	41	19	25	33
$0.4 < \kappa \leq 0.6$	7	2	5	4
$0.6 < \kappa \leq 0.8$	2	0	1	0
$0.8 < \kappa \leq 1.0$	1	1	1	1

Table 6: The number of pairs of native annotators with the IAA scores (Cohen's κ) in each of five score ranges, for each of the four types of sentences (or text snippets).

κ range	Original	Gramm.	Split	Addit.
$0 < \kappa \leq 0.2$	153	175	160	154
$0.2 < \kappa \leq 0.4$	34	13	22	28
$0.4 < \kappa \leq 0.6$	2	1	6	7
$0.6 < \kappa \leq 0.8$	1	1	2	1
$0.8 < \kappa \leq 1.0$	0	0	0	0

Table 7: The number of pairs of non-native annotators with the IAA scores (Cohen's κ) in each of the five score ranges, for each of the four types of sentences (or text snippets).

In this set of experiments, we calculate the unweighted Cohen's kappa (κ) as a measure of inter-annotator agreement among each pair of annotators within the group of native annotators, and within the group of non-native annotators (a total of 190 pairs in each group).[2] At the same time, we control for the type of evaluation (absolute vs. relative), and for the sentence type (or, in the case of the relative simplicity evaluation, for the sentence pair type).

7.1. IAA in Absolute Evaluations

In the case of the absolute simplicity evaluation, each sentence was marked with a 1–5 score. For each group (native or non-native) of 190 annotator pairs, and for each of the five IAA score ranges, we count the number of annotator pairs with the IAA in that range (Tables 6 and 7).

The obtained results confirm the earlier findings that native speakers have better IAA than non-native speakers (Yatskar et al., 2010; Yimam et al., 2017). The most striking, however, is that in both annotator groups, over 73% of annotator pairs (regardless of the annotators group and the sentence type) have a very low IAA agreement ($0 < \kappa \leq 0.2$).

Both annotator groups show similar trends in how sentence type influences the IAA (see Figures 1 and 2). In both cases, the lowest IAA is achieved for the grammatical sentences and those that were split.

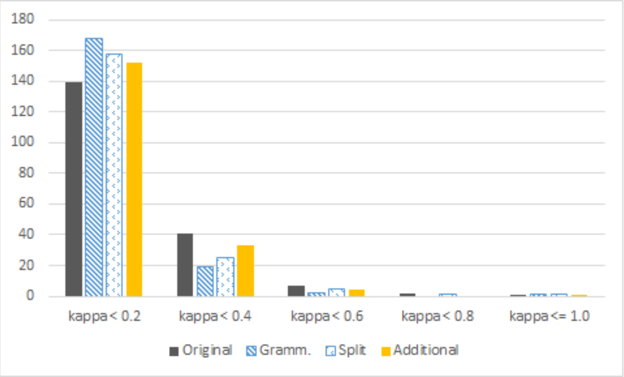

Figure 1: Distribution of the IAA scores depending on the sentence type for the native annotators.

[2] Although the IAA can be calculated for multiple annotators, in TS evaluation, it is a common practice to report the average pairwise IAA calculated as the (un)weighted Cohen's kappa.

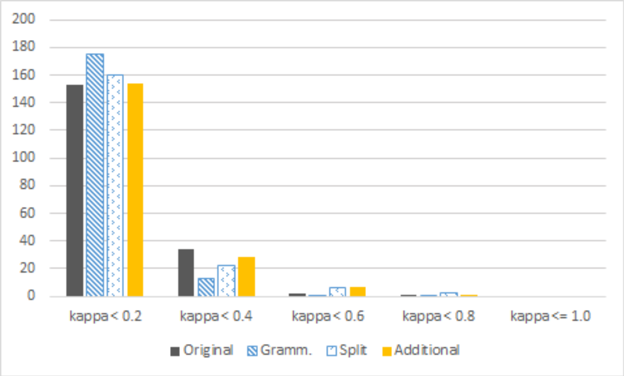

Figure 2: Distribution of the IAA scores depending on the sentence type for the non-native annotators.

κ range	O>G	O>S	O>A	G>S	G>A	S>A
$0 < \kappa \leq 0.2$	133	118	23	128	125	149
$0.2 < \kappa \leq 0.4$	40	49	0	40	49	21
$0.4 < \kappa \leq 0.6$	14	18	40	17	12	18
$0.6 < \kappa \leq 0.8$	1	3	70	4	1	0
$0.8 < \kappa \leq 1.0$	2	2	57	1	3	2

Table 8: The number of pairs of native annotators with the IAA scores (Cohen's κ) in each of the five score ranges, for each of the four types of sentences (or text snippets).

7.2. IAA in Relative Evaluations

In the case of the relative evaluation of simplicity, the annotators are asked to answer 'yes'/'no' to the question whether one sentence is simpler than the other. For each group (native or non-native) of 190 annotator pairs, and each of the five IAA score ranges, we count the number of annotator pairs with the IAA in that range (Tables 8 and 9). As expected, the IAA is better for the relative evaluations than for the absolute evaluations within both groups of annotators. We found significantly higher number of annotator pairs with the κ between 0.4 and 0.6 in the relative evaluations than in the absolute evaluations. The most striking is, however, that only the relative evaluations that compare the original sentences with the fully simplified sentences (O>A) achieve a very high number of good IAA scores, within both annotator groups (with better inter-annotator agreements within the native than within the non-native annotators).

These results call for caution in current TS evaluation methods, where human evaluations are either crowdsourced or where the IAA is not checked (those studies that use very few annotators usually have more experienced and well trained annotators, with a high IAA reported).

κ range	O>G	O>S	O>A	G>S	G>A	S>A
$0 < \kappa \leq 0.2$	120	142	59	141	152	148
$0.2 < \kappa \leq 0.4$	66	43	6	43	34	40
$0.4 < \kappa \leq 0.6$	4	5	40	5	4	2
$0.6 < \kappa \leq 0.8$	0	0	54	1	0	0
$0.8 < \kappa \leq 1.0$	0	0	31	0	0	0

Table 9: The number of pairs of non-native annotators with the IAA scores (Cohen's κ) in each of five score ranges, for each of the four types of sentences (or text snippets).

Additionally, these results show that we can have reliable relative comparisons of simplicity only in those cases where the differences between the two sentences are obvious (as in the case of the original sentence being compared with the fully simplified sentence).

8. Conclusions

In this study, we explored the differences in how native and non-native speakers perceive sentence simplicity, aiming for better understanding of their simplification needs and for better understanding how the choice of evaluators (native or non-native speakers) can influence the results of simplicity assessment. To do so, we built a new dataset with human evaluation of simplicity of text snippets, carefully controlling for various factors that could influence the perceived simplicity.

We found that native and non-native annotators differently reward both grammaticality and various simplification operations in their simplicity scores, and that grammaticality influences simplicity score more than any other (non-lexical) simplification transformation. These results imply that we should be cautions when mixing native and non-native annotators in TS evaluation, or when we use native annotators for evaluating the simplicity of the sentences simplified for non-native speakers, and vice versa.

The presented results also show that native and non-native speakers have different needs for a better understanding of the instructions in the IT troubleshooting domain. Grammatical correctness of the sentences influences the perceived simplicity of the sentences more in native than in non-native speakers. After the sentences have been split, additional simplification operations (removing superfluous words, conversion of passive to active voice, disambiguation of meaning, and conversion to the canonical subject-verb-object form) only improves the simplicity for native speakers.

We further found that inter-annotator agreements measured as the Cohen's kappa (κ) are, in most cases, very low for the majority of annotator pairs, regardless of the annotators group (native vs. non-native) and the type of evaluation performed (absolute evaluation of sentences on a 1–5 level scale vs. relative evaluation with a 'yes'/'no' answer to the question whether one sentence is simpler than the other). The only reliable results (in terms of the Cohen's kappa) seem to be those for the relative comparison of the original and fully simplified sentences.

These results call for special caution in TS evaluation, showing that crowdsourced evaluation without checking the inter-annotator agreements can result in misleading results. They also show that the currently used absolute evaluation on a 1–5 Likert scale might not be suitable for the sentence simplicity assessment, as it leads to high differences in scores across annotators even if they all belong to the same annotators group (native or non-native).

9. Acknowledgements

We thank Nikitha Sukumaran for helping in data collection and evaluation.

10. Bibliographical References

Aluísio, S. M. and Gasperin, C. (2010). Fostering digital inclusion and accessibility: The porsimples project for simplification of portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, YIWCALA '10, pages 46–53, Stroudsburg, PA, USA. Association for Computational Linguistics.

Angrosh, M., Nomoto, T., and Siddharthan, A. (2014). Lexico-syntactic text simplification and compression with typed dependencies. In *Procedings of COLING 2014*, pages 1996–2006.

Aranzabe, M. J., Díaz De Ilarraza, A., and González, I. (2012). First Approach to Automatic Text Simplification in Basque. In *Proceedings of the first Natural Language Processing for Improving Textual Accessibility Workshop (NLP4ITA)*.

Baeza-Yates, R., Rello, L., and Dembowski, J. (2015). Cassa: A context-aware synonym simplification algorithm. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1385. ACL.

Barlacchi, G. and Tonelli, S. (2013). ERNESTA: A sentence simplification tool for childrens stories in italian. In *Computational Linguistics and Intelligent Text Processing*.

Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S., and Tait, J. (1999). Simplifying text for language-impaired readers. In *Proceedings of the 9th Conference of the European Chapter of the ACL (EACL'99)*, pages 269–270.

Coster, W. and Kauchak, D. (2011a). Learning to Simplify Sentences Using Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1–9.

Coster, W. and Kauchak, D. (2011b). Simple English Wikipedia: a new text simplification task. In *Proceedings of ACL&HLT*, pages 665–669.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram coocurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

Drndarević, B., Štajner, S., Bott, S., Bautista, S., and Saggion, H. (2013). Automatic Text Simplication in Spanish: A Comparative Evaluation of Complementing Components. In *Proceedings of CICLing 2012*, Lecture Notes in Computer Science, pages 488–500. Springer.

Freyhoff, G., Hess, G., Kerr, L., Tronbacke, B., and Van Der Veken, K., (1998). *Make it Simple, European Guidelines for the Production of Easy-toRead Information for People with Learning Disability*. ILSMH European Association, Brussels.

Glavaš, G. and Štajner, S. (2013). Event-Centered Simplication of News Stories. In *Proceedings of the Student Workshop at RANLP 2013*, pages 71–78.

Glavaš, G. and Štajner, S. (2015). Simplifying Lexical Simplification: Do We Need Simplified Corpora? In *Proceedings of the ACL&IJCNLP 2015 (Volume 2: Short Papers)*, pages 63–68.

Horn, C., Manduca, C., and Kauchak, D. (2014). Learning a lexical simplifier using wikipedia. In *Proceedings of ACL 2014 (Short Papers)*, pages 458–463.

Hwang, W., Hajishirzi, H., Ostendorf, M., and Wu, W. (2015). Aligning Sentences from Standard Wikipedia to Simple Wikipedia. In *Proceedings of NAACL&HLT*, pages 211–217.

Mencap, (2002). *Am I making myself clear? Mencap's guidelines for accessible writing*.

Nisioi, S., Štajner, S., Ponzetto, S. P., and Dinu, L. P. (2017). Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 85–91.

Orasan, C., Evans, R., and Dornescu, I., (2013). *Towards Multilingual Europe 2020: A Romanian Perspective*, chapter Text Simplification for People with Autistic Spectrum Disorders, pages 287–312. Romanian Academy Publishing House, Bucharest.

Paetzold, G. H. and Specia, L. (2016a). Benchmarking lexical simplification systems. In *Proceedings of LREC*, pages 3074–3080.

Paetzold, G. H. and Specia, L. (2016b). Unsupervised lexical simplification for non-native speakers. In *Proceedings of the 30th AAAI*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.

PlainLanguage. (2011). Federal plain language guidelines.

Rello, L., Baeza-Yates, R., Bott, S., and Saggion, H. (2013). Simplify or help? Text simplification strategies for people with dyslexia. In *Proceedings of W4A conference*.

Saggion, H., Štajner, S., Bott, S., Mille, S., Rello, L., and Drndarevic, B. (2015). Making It Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. *ACM Transactions on Accessible Computing*, 6(4):14:1–14:36.

Siddharthan, A. and Angrosh, M. A. (2014). Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 722–731.

Specia, L. (2010). Translating from complex to simplified sentences. In *Proceedings of the 9th international conference on Computational Processing of the Portuguese Language (PROPOR)*, volume 6001 of *Lecture Notes in Computer Science*, pages 30–39. Springer Berlin Heidelberg.

Štajner, S. and Glavaš, G. (2017). Leveraging event-based semantics for automated text simplification. *Expert Systems With Applications, Elsevier*, 82:383–395.

Štajner, S. and Nisioi, S. (2018). A Detailed Evaluation of Neural Sequence-to-Sequence Models for In-domain and Cross-domain Text Simplification. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*.

Štajner, S., Mitkov, R., and Saggion, H. (2014). One Step Closer to Automatic Evaluation of Text Simplification Systems. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) at EACL*.

Štajner, S., Bechara, H., and Saggion, H. (2015a). A Deeper Exploration of the Standard PB-SMT Approach to Text Simplification and its Evaluation. In *Proceedings of ACL&IJCNLP (Volume 2: Short Papers)*, pages 823–828.

Štajner, S., Calixto, I., and Saggion, H. (2015b). Automatic Text Simplification for Spanish: Comparative Evaluation of Various Simplification Strategies. In *Proceedings of RANLP 2015*, pages 618–626.

Štajner, S., Popović, M., Saggion, H., Specia, L., and Fishel, M. (2016). Shared Task on Quality Assessment for Text Simplification. In *Proceedings of the LREC Workshop on Quality Assessment for Text Simplification (QATS)*, pages 22–31.

Woodsend, K. and Lapata, M. (2011). Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 409–420.

Wubben, S., van den Bosch, A., and Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL): Long Papers - Volume 1*, pages 1015–1024. Association for Computational Linguistics.

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

Yatskar, M., Pang, B., Danescu-Niculescu-Mizil, C., and Lee, L. (2010). For the sake of simplicity: unsupervised extraction of lexical simplifications from wikipedia. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 365–368, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yimam, S. M., Štajner, S., Riedl, M., and Biemann, C. (2017). Multilingual and Cross-Lingual Complex Word Identification. In *Proceedings of RANLP*, pages 813–822, Varna, Bulgaria.

Zhang, X. and Lapata, M. (2017). Sentence Simplification with Deep Reinforcement Learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594.