The C-Score – Proposing a Reading Comprehension Metrics as a Common Evaluation Measure for Text Simplification

Irina Temnikova
Linguistic Modelling Department,
Institute of Information
and Communication Technologies,
Bulgarian Academy of Sciences
irina.temnikova@gmail.com

Galina Maneva
Lab. of Particle and Astroparticle Physics,
Institute of Nuclear Research
and Nuclear Energy,
Bulgarian Academy of Sciences
galina.maneva@gmail.com

Abstract

This article addresses the lack of common approaches for text simplification evaluation, by presenting the first attempt for a common evaluation metrics. The article proposes reading comprehension evaluation as a method for evaluating the results of Text Simplification (TS). An experiment, as an example application of the evaluation method, as well as three formulae to quantify reading comprehension, are presented. The formulae produce an unique score, the C-score, which gives an estimation of user's reading comprehension of a certain text. The score can be used to evaluate the performance of a text simplification engine on pairs of complex and simplified texts, or to compare the performances of different TS methods using the same texts. The approach can be particularly useful for the modern crowdsourcing approaches, such as those employing the Amazon's Mechanical Turk[1] or CrowdFlower[2]. The aim of this paper is thus to propose an evaluation approach and to motivate the TS community to start a relevant discussion, in order to come up with a common evaluation metrics for this task.

1 Context and Motivation

Currently, the area of Text Simplification (TS) is getting more and more attention. Starting as early as in the 1996, Chandrasekar et al. proposed an approach for TS as a pre-processing step before feeding the text to a parser. Next, the PSET project (Devlin, 1999; Canning, 2002), proposed two modules for simplifying text for aphasic readers. The text simplification approaches continued in 2003 with Siddharthan (2003) and Inui et al. (2003), and through the 2005-2006 until the recent explosion of TS approaches in 2010-2012. Recently, several TS-related workshops took place: PITR 2012 (Williams et al., 2012), SLPAT 2012 (Alexandersson et al., 2012), and NLP4ITA 2012[3] and 2013. As in confirmation with the text simplification definition as the "process for reducing text complexity at different levels" (Temnikova, 2012), the TS approaches tackle a variety of text complexity aspects, ranging from lexical (Devlin, 1999; Inui et al., 2003; Elhadad, 2006; Gasperin et al., 2009; Yatskar et al., 2010; Coster and Kauchak, 2011; Bott et al., 2012; Specia et al., 2012; Rello et al., 2013; Drndarević et al., 2013), syntactic (Chandrasekar et al., 1996; Canning, 2002; Siddharthan, 2003; Inui et al., 2003; Gasperin et al., 2009; Zhu et al., 2010; Woodsend and Lapata, 2011; Coster and Kauchak, 2011; Drndarević et al., 2013), to discourse/cohesion (Siddharthan, 2003). The variety of problems tackled by the TS approaches differ, according to their final aim: (1) being a pre-processing step of an input to text processing applications, or (2) addressing the reading difficulties of specific groups of readers. The first type of final application ranges between parser input (Chandrasekar et al., 1996), small screens displays (Daelemans et al., 2004; Grefenstette, 1998), text summarization (Vanderwende et al., 2007), text extraction (Klebanov et al., 2004), semantic role labeling (Vickrey and Koller, 2008) and Machine Translation (MT) (Ruffino, 1982; Streiff, 1985).The TS approaches addressing specific human reading needs, instead, address readers with low levels of literacy (Siddharthan, 2003;

[1] http://aws.amazon.com/mturk/. Last accessed on May 3rd, 2013.

[2] http://crowdflower.com/. Last accessed on June 14th, 2013.

[3] http://www.taln.upf.edu/nlp4ita/. Last accessed on May 3rd, 2013.

Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations, pages 20–29,
Sofia, Bulgaria, August 4-9 2013.

Gasperin et al., 2009; Elhadad, 2006; Williams and Reiter, 2008), language learners (Petersen and Ostendorf, 2007), and readers with specific cognitive and language disabilities. The TS approaches, addressing this last type of readers target those suffering from aphasia (Devlin, 1999; Canning, 2002), deaf readers (Inui et al., 2003), dyslexics (Rello et al., 2013) and the readers with general disabilities (Max, 2006; Drndarević et al., 2013).

Despite the large number of current work in TS, there has been almost no attention to defining common text simplification evaluation approaches, which would allow the comparison of different TS systems. Until the present moment, usually, each approach has applied his/her own methods and materials, often taken from other Natural Language Processing (NLP) fields, making the comparison difficult or impossible.

The aim of this paper is thus to propose an evaluation method and to foster the discussion of this topic in the text simplification community, as well as to motivate the TS community to come up with common evaluation metrics for this task.

Next, Section 2 will describe the existing approaches to evaluating TS, as well as the few attempts towards offering a common evaluation strategy. After that, the next sections will present our evaluation approach, starting with Section 3 describing its context, Section 4 presenting the formulae, Section 5 offering the results, and finally Section 6, providing a Discussion and the Conclusions.

2 Evaluation Methods in Text Simplification

As mentioned in the previous section, until now, the different authors adopted different combinations of metrics, without reaching to a common approach, which would allow the comparison of different systems. As the different TS evaluation methods are applied on a variety of different text units (words, sentences, texts), this makes the comparison between approaches even harder. As the aim of this article is to propose a text simplification evaluation metrics which would take into account text comprehensibility and reading comprehension, in this discussion we will focus mostly on the approaches, whose aim is to simplify texts for target readers and their evaluation strategies.

The existing TS evaluation approaches focus either on the quality of the generated text/sentences, or on the effectiveness of text simplification on reading comprehension. The first group of approaches include human judges ratings of simplification, content preservation, and grammaticality, standard MT evaluation scores (BLEU and NIST), a variety of other automatic metrics (perplexity, precision/recall/F-measure, and edit distance). The methods, aiming to evaluate the text simplification impact on reading comprehension, use, instead, reading speed, reading errors, speech errors, comprehension questions, answer correctness, and users' feedback. Several approaches use a variety of readability formulae (the Flesch, Flesch-Kincaid, Coleman-Liau, and Lorge formulae for English, as well as readability formulae for other languages, such as for Spanish). Due to the criticisms of readability formulae (DuBay, 2004), which often restrict themselves to a very superficial text level, they can be considered to stand on the borderline between the two previously described groups of TS evaluation approaches. As can be seen from the discussion below, different TS systems employ a combination of the listed evaluation approaches.

As one of the first text simplification systems for target reader populations, PSET, seems to have applied different evaluation strategies for different of its components, without running an evaluation of the system as a whole. The lexical simplification component (Devlin, 1999), which replaced technical terms with more frequent synonyms, was evaluated via user feedback, comprehension questions and the use of the Lorge readability formula (Lorge, 1948). The syntactic simplification system evaluated its single components and the system as a whole from different points of view, to a different extent, and used different evaluation strategies. Namely, the text comprehensibility was evaluated via reading time and answers' correctness given by sixteen aphasic readers; the components replacing passive with active voice and splitting sentences were evaluated for content preservation and grammaticality via four human judges' ratings; and finally, the anaphora resolution component was evaluated using precision and recall. Siddharthan (2003) did not carry out evaluation with target readers, while three human judges rated the grammaticality and the meaning preservation of ninety-five sentences. Gasperin et al. (2009) used precision, recall and f-measure. Other approaches, using human judges are those of Elhadad (2006),

who also used precision and recall and Yatskar et al. (2010), who employed three annotators comparing pairs of words and indicating them same, simpler, or more complex. Williams and Reiter (2008) run two experiments, the larger one involving 230 subjects and measured oral reading rate, oral reading errors, response correctness to comprehension questions and finally, speech errors. Drndarevic et al. (2013) used 7 readability measures for Spanish to evaluate the degree of simplification, and twenty-five human annotators to evaluate on a Likert scale the grammaticality of the output and the preservation of the original meaning. The recent approaches considering TS as an MT task, such as Specia (2010), Zhu et al. (2010), Coster and Kauchak (2011) and Woodsend and Lapata (2011), apply standard MT evaluation techniques, such as BLEU (Papineni et al., 2002), NIST (Doddington, 2002), and TERp (Snover et al., 2009). In addition, Woodsend and Lapata (2011) apply two readability measures (Flesch-Kincaid, Coleman-Liau) to evaluate the actual reduction in complexity and human judges ratings for simplification, meaning preservation, and grammaticality. Zhu et al. (2010) apply the Flesch readability score (Flesch, 1948) and n-gram language model perplexity, and Coster and Kauchak (2011) – two additional automatic techniques (the word-level-F1 and simple string accuracy), taken from sentence compression evaluation (Clarke and Lapata, 2006).

As we consider that the aim of text simplification for human readers is to improve text comprehensibility, we argue that reading comprehension must be evaluated, and that evaluating just the quality of produced sentences is not enough. Differently from the approaches that employ human judges, we consider that it is better to test real human comprehension with target readers populations, rather than to make conclusions about the extent of population's understanding on the basis of the opinion of a small number of human judges. In addition, we consider that measuring reading speed, rate, as well as reading and speed errors, requires much more complicated and expensive tools, than having an online system to measure time to reply and recognize correct answers. Finally, we consider that cloze tests are an evaluation method that cannot really reflect the complexity of reading comprehension (for example for measuring manipulations of the syntactic structure of sentences), and for this reason, we select multiple-choice questions as the testing method, which we consider the most reflecting the specificities of the complexity of a text, more accessible than eye-tracking technologies, and more objective than users' feedback. The approach does not explicitly evaluate the fluency, grammaticality and content preservation of the simplified text, but can be coupled with such additional evaluation.

The closest to ours approach is that of Rello et al. (2013) who evaluated reading comprehension with over ninety readers with and without dyslexia. Besides using eye-tracking (reading time and fixations duration), different reading devices, and users rating the text according to how easy it is it read, to understand and to remember, they obtain also a comprehension score based on multiple-choice questions (MCQ) with 3 answers (1 correct, 1 partially correct and 1 wrong). The difference with our approach is that we consider that having only one correct answer (as suggested by Gronlund (1982)), is a more objective evaluation, rather than having one partially correct answer, which would introduce subjectivity in evaluation.

To support our motivation, some state-of-the-art approaches state the scarcity of evaluation with target readers (Williams and Reiter, 2008), note that there are no commonly accepted evaluation measures (Coster and Kauchak, 2011), attempt to address the need of developing reading comprehension evaluation methods (Siddharthan and Katsos, 2012), and propose common evaluation frameworks (Specia et al., 2012; De Belder and Moens, 2012). More concretely, Siddhathan and Katsos (2012) propose the magnitude estimation of readability judgements and the delayed sentence recall as reading comprehension evaluation methods. Specia et al. (2012) provide a lexical simplification evaluation framework in the context of Semeval-2012. The evaluation is performed using a measure of inter-annotator agreement, based on Cohen (1960). Similarly, De Belder and Moens (2012) propose a dataset for evaluating lexical simplification. No common evaluation framework has been yet developed for syntactic simplification.

As seen in the overview, besides the multitude of existing approaches, and the few approaches attempting to propose a common evaluation framework, there are no widely accepted evaluation metrics or methods, which would allow the com-

parison of existing approaches. The next section presents our evaluation approach, which we offer as a candidate for common evaluation metrics.

3 Proposed Evaluation Metrics

3.1 The Evaluation Experiment

The metrics proposed in this article, was developed in the context of a previously conducted large-scale text simplification evaluation experiment (Temnikova, 2012). The experiment aimed to determine whether a manual, rule-based text simplification approach (namely a controlled language), can re-write existing texts into more understandable versions. Impact on reading comprehension was necessary to evaluate, as the purpose of text simplification was to enhance in first place the reading comprehension of emergency instructions. The controlled language used for simplification was the Controlled Language for Crisis Management (CLCM, more details in (Temnikova, 2012)), which was developed on the basis of existing psychological and psycholinguistic literature discussing human comprehension under stress, which ensures its psychological validity. The text units evaluated in this experiments were whole texts, and more concretely pairs of original texts and their simplified versions. We argue that using whole texts for measuring reading comprehension is better than single sentences, as the texts provide more context for understanding. The experiment took place in the format of an online experiment, conducted via a specially developed web interface, and required users to read several texts and answer Multiple-Choice Questions (MCQ), testing the readers' understanding of each of the texts. Due to the purpose of the text simplification (emergency situations simulation), users were required to read the texts in a limited time, as to imitate a stressful situation with no time to think and re-read the text. This aspect will not be taken into account in the evaluation, as the purpose is to propose a general formula, applicable to a variety of different text simplification experiments. After reading the text in a limited time, the text was hidden from the readers, and they were presented with a screen, asking if they were ready to proceed with the questions. Next, each question was displayed one by one, along with its answers, with the readers not having the option to go back to the text. In order to ensure the constant attention of the readers and to reduce readers' tiredness fact or, the texts were kept short (about 150-170 words each), and the number of texts to be read by the reader was kept to four. In addition, to ensure comparability, all the texts were selected in a way to be more or less of the same length. The experiment employed a collection of a total of eight texts, four of which original, non simplified ('complex') versions, and the other four – their manually simplified versions. Each user had to read two complex and two simplified texts, none of which was a variant of the other. The interface automatically randomized the order of displaying the texts, to ensure that different users would get different combinations of texts in one of the following two different sequences:

- Complex-Simplified-Complex-Simplified
- Simplified-Complex-Simplified-Complex

This was done in order to minimize the impact of the order of displaying the texts on the text comprehension results. After reading each text, the readers were prompted to answer between four and five questions about each text. The MCQ method was selected as it is considered being the most objective and easily measurable way of assessing comprehension (Gronlund, 1982). The number of questions and answers was selected in a way to not tire the reader (four to five questions per text and four to five answers for each question), and the questions and answers themselves were designed following the the best MCQ practices (Gronlund, 1982). Some of the practices followed involved ensuring that there is only one correct answer per question, making all wrong answers (or 'distractors') grammatically, and as text length consistent with the correct answer, in order to avoid giving hints to the reader, and making all distractors plausible and equally attractive. Similarly to the texts, the questions and answers were also displayed in different order to different readers, to avoid that the order influences the comprehension results. The correct answer was displayed in different positions to avoid learning its position and internally marked in a way to distinguish it during evaluation from all the distractors in whatever position it was displayed. The questions required understanding of key aspects of the texts, to avoid relying on pure texts' memorization (such as under which conditions what was supposed to be done, explanations, and the order in which actions needed to be taken). The information, evaluating

the users' comprehension, collected during the experiment, was, on one hand the time for answering each question, and on the other hand, the number of correct answers given by all participants while replying to the same question. Besides the fact that we used a specially developed interface, this evaluation approach can be applied to any experiment employing an interface capable of calculating the time for answering and to distinguish the correct answers from the incorrect ones.

The efficiency of the experiment design was thoroughly tested by running it through several rounds of pilot experiments and requiring participants' feedback.

We claim that the evaluation approach proposed in this paper can be applied to more simply organized experiments, as the randomization aspects are not reflected in the evaluation formulae.

The final experiment involved 103 participants, collected via a request sent to several mailing lists. The participants were 55 percent women and 44 percent male, and ranged from undergraduate students to retired academicians (i.e. corresponded to nineteen to fifty-nine years old). As the experiment allowed entering lots of personal data, it was also known that participants had a variety of professions (including NLP people, teachers, and lawyers), knew English from the beginner through intermediate, to native level, and spoke a large variety of native languages, allowing to have native speakers from many of the World's language families (Non Indo-European and Indo-European included). Figure 1 shows the coarse-grained classification made at the time of the experiment, and the distribution of participants per native languages. A subset of specific native language participants will be selected to give an example of applying the evaluation metrics to a real evaluation experiment.

In order to obtain results, we have asked the participants to enter a rich selection of information, and recorded the chosen answer (be it correct or not), and the time which each participant employed to give each answer (correct or wrong). Table 1 shows the data we recorded for each single answer of every participant.

The data in Table 1 is: *Entry id* is each given answer, the *Domain background* (answer *y* – yes and *n* – no) indicates whether the participant has any previous knowledge of the experiment (crisis management) domain. As each text, question and complex/simplified texts pair are given reference numbers, respectively *Text number*, *Question number*, and *Texts pair number* record that. As required by the evaluation method, each entry records also the *Time to reply* each question (measured in 'milliseconds'), and the *Answer number*. As said before, the correct answers are marked in a special way, allowing to distinguish them at a later stage, when counting the number of correct answers.

Type	**Example**
Entry id	1
Age of the participant	24
Gender of the participant	f
Profession of the participant	Student
Domain background (y/n)	n
Native lang.	English
Level of English	Native
Text number	4
Exper. completed (0/1)	1
User number	1
Question number	30
Answer number	0
Time to reply	18695
Texts pair number	1

Table 1: Participant's information recorded for each answer.

3.2 Definitions and Evaluation Hypotheses

In order to correctly evaluate the performance of the text simplification method on the basis of the above described experiment, the data obtained was thoughtfully analyzed. The two criteria selected to best describe the users' performance were time to reply and number of correct answers. The evaluation was done offline, after collecting the data from the participants. The evaluation analysis aimed to test the following **two hypotheses**:

If the text simplification approach has a positive impact on the reading comprehension:

1. The percentage of correct answers given for the simplified text will be higher than the percentage of correct answers given for the complex text.

2. The time to recognize the correct answer and reply correctly to the questions about the simplified text will be significantly lower than the time to recognize the correct answer and

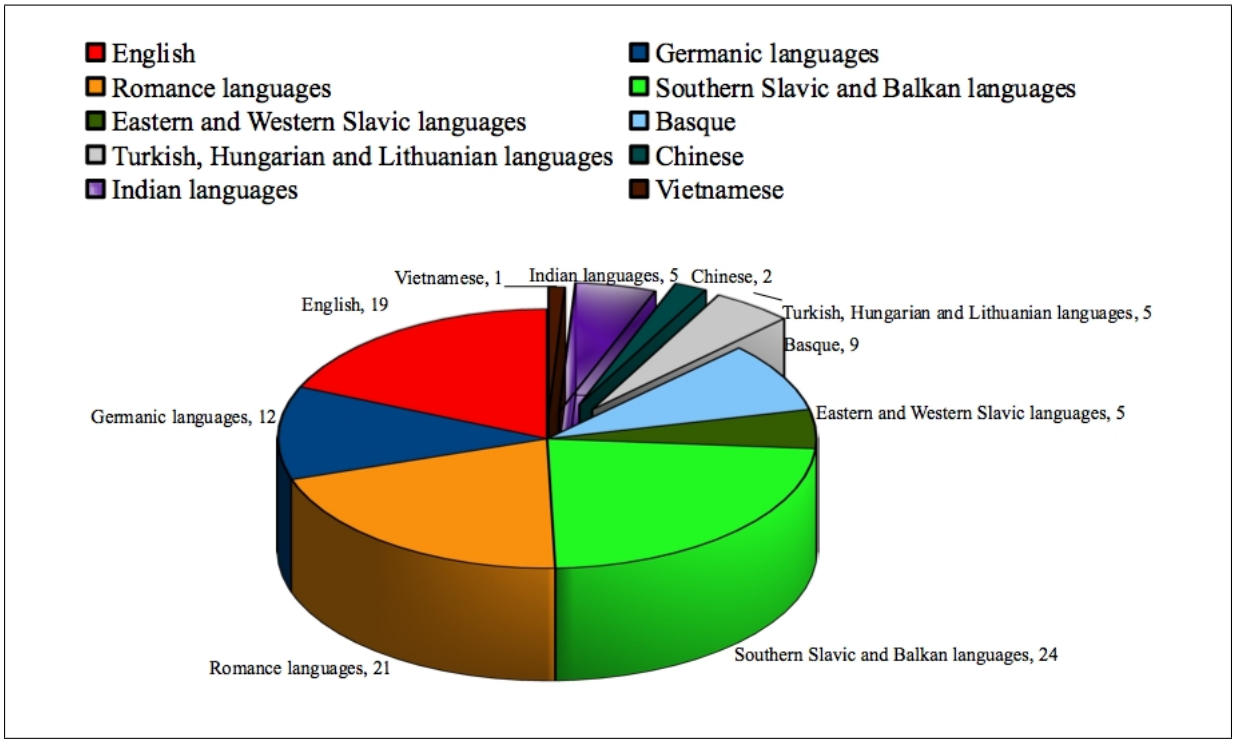

Figure 1: Coarse-grained distribution of participants per native languages.

reply correctly to the questions about the complex text.

The two hypotheses were tested previously by employing only the key variables (time to reply and number of correct answers). It has been proven that comprehension increases with the percentage of correct answers and decreases with the increase of the time to reply. On the basis of these facts, we define the **C-Score** (a text Comprehension Score) – an objective evaluation metrics, which allows to give a reading comprehension estimate to a text, or to compare two texts or two or more text simplification approaches. The C-Score is calculated text per text. In order to address a variety of situations, we propose three versions of the C-Score, which cover, gradually, all possible variables which can affect comprehension in such an experiment. In the following sections we present their formulae, the variables involved, and discuss their results, advantages and shortcomings.

3.3 The C-Score Version One. The C-Score Simple.

Given a text comprehension experiment featuring *n* texts with *m* questions with *r* answers each, an ability to measure time to reply to questions and to recognize the correct answers, we define the *C-Score Simple* as given below:

$$C_{simple} = \frac{Pr}{t_{mean}} \quad (1)$$

Where: *Pr* is the percentage of correct answers, from all answers given to all the questions about this text, and *t* is the average time to reply to all questions about this text (both with a correct and a wrong answer). The time is expressed in arbitrary seconds-based units, depending on the experiment. The logic behind this formula is simple: we consider that comprehension increases with the percentage of correctly answered questions, and diminishes if the mean time to answer questions increases.

3.4 The C-Score Version Two. C-Score Complete.

The C-Score complete takes into consideration a rich selection of variables reflecting the questions and answers complexity. In this C-Score version, we consider that the experiment designers will select short texts (e.g. 150 words) of a similar length, with the aim to reduce participants' tiredness factor, as we did in our experimental settings.

$$C_{complete} = \frac{Pr}{Nq} \sum_{q=1}^{Nq} \frac{Qs(q)}{t_{mean}(q)} \quad (2)$$

In this formula, *Pr* is the percentage of correct answers by all participants for this text, *Nq* is the

number of questions of this text (4-5 in our experiment), and *t* is the average time to reply to all questions about this text (4-5 in our experiment). We introduce the concept Question Size, (*Qs*), which is calculated for each question and takes into account the number of answers of the question (*Na*), the question length in words (*Lq*), and the total length in words of its answers (*La*):

$$Qs = Na(Lq + La) \tag{3}$$

We consider that the number of questions negatively influences the comprehension results, as the reader gets cognitively tired to process more and more questions about different key aspects of the text. In addition, Gronlund (1982) suggests to restrict the number of questions per text to four-five to achieve better learning. For this reason, we consider that comprehension decreases, if the number of questions is higher. We also consider that answering correctly/faster to a difficult question shows better text comprehension than giving fast a correct answer to a simply-worded question. For this reason we award question difficulty, and we place it above the fraction.

3.5 The C-Score Version Three. C-Score Textsize.

Finally, the last version of C-Score takes into account the case when the texts used for comparison can be of a different length, and in this way, the texts' complexity (for example, when comparing the results of two different TS engines, without having access to the same texts). For this reason, the C-Score 3 considers the text length (called *text size, Ts*) of the texts used in the experiment. As a longer text will be more difficult to understand than a shorter text, the text length is placed near the *percentage of correct answers*.

$$C_{textsize} = \frac{PrTs}{Nq} \sum_{q=1}^{Nq} \frac{Qs(q)}{t_{mean}(q)} \tag{4}$$

4 C-Score Results

We have implemented and applied the above described formulae to the experimental data, presented in Section 3.1. As we have only one text simplification approach, two user scenarios are presented:

1. Original ('Complex') vs. Simplified ('Simple') pairs of texts comparison. The subset of participants are the speakers of Basque, Turkish, Hungarian, Lithuanian, Vietnamese, Chinese, and Indian languages. All three formulae have been applied.

2. Comparison of the comprehension of the same text of readers from different subgroups. The readers have been divided by age. This scenario can be used to infer psycho-linguistic findings about the reading abilities of different participants.

Please note that the texts pairs are: Text 1 and 2; Text 3 and 4; Text 5 and 6; and Text 7 and 8. In each couple, the first text is complex and the second is its simplified version. The results for the first evaluation scenario are respectively displayed in Table 2 for *C-Score Simple*, Table 3 for *C-Score Complete* and Table 4 for *C-Score Textsize*. The results of C-Score Complete have been multiplied per 100 for better readability. As a reminder, we consider that higher the score is, better is text comprehension. From this point of view, if the text simplification approach was successful, Text 2 (Simplified) should have a higher C-Score than its original, complex Text 1, Text 4 (Simplified) should have a higher C-Score than its original Text 3, Text 6 (Simplified) – a higher score than the complex Text 5, and Text 8 (Simplified) – a higher score than its original Text 7.

In the second scenario, the participants data has been divided into data relevant to participants under 45 years old (ninety-two participants) and into participants over 45 years old (eleven participants). In this case only the C-Score Simple has been applied. The results of this evaluation are shown in Table 5. As our aim is to compare the reading abilities of different ages of people, and not the results of text simplification, only the complex texts are taken into account. The results show that the comprehension score of participants under 45 years old is higher for all texts (despite the uneven participants' distribution), except in the case of complex Text 5.

A similar phenomenon can be observed in Tables 2, 3 and 4, where in all text pairs, except for pair 3, i.e. Texts 5 and 6 (where can be observed the opposite), the simplified text has a higher comprehension score than its complex original. The hypothesis about the different behavior of Text 5 and 6 is that it is text-specific. This is confirmed by Table 5, which shows that besides the big dif-

Text number	C-Score Simple
Text 1 (Complex)	21.3
Text 2 (Simplified)	35.3
Text 3 (Complex)	24.8
Text 4 (Simplified)	34.9
Text 5 (Complex)	36.8
Text 6 (Simplified)	23.6
Text 7 (Complex)	40.5
Text 8 (Simplified)	51.5

Table 2: Experiment results for C-Score Simple.

ferences in reading comprehension between participants under 45 years old and participants over 45 years old, Text 5 has more or less the same comprehension score for both groups of readers. From this fact we can assume that this text is probably fairly easy, so this type of combination of text simplification rules does not simplify it, and instead, when applied makes it less comprehensible or more awkward for the human readers.

Text number	C-Score Complete
Text 1 (Complex)	66.3
Text 2 (Simplified)	114.3
Text 3 (Complex)	65.3
Text 4 (Simplified)	89.9
Text 5 (Complex)	104.0
Text 6 (Simplified)	66.9
Text 7 (Complex)	106.7
Text 8 (Simplified)	153.0

Table 3: Experiment results for C-Score Complete.

Text number	C-Score Textsize
Text 1 (Complex)	109.5
Text 2 (Simplified)	192.0
Text 3 (Complex)	107.7
Text 4 (Simplified)	131.3
Text 5 (Complex)	171.6
Text 6 (Simplified)	102.4
Text 7 (Complex)	176.1
Text 8 (Simplified)	263.3

Table 4: Experiment results for C-ScoreTextsize.

5 Discussion and Conclusions

This article has presented an extended discussion of the methods employed for evaluation in the text

Text number	Under 45	Over 45
Text 1 (Complex)	39.7	22.5
Text 3 (Complex)	37.2	18.4
Text 5 (Complex)	38.4	38.9
Text 7 (Complex)	54.3	35.9

Table 5: C-Score Simple for one text.

simplification domain. In order to address the lack of common or standard evaluation approaches, this article proposed three evaluation formulae, which measure the reading comprehension of produced texts. The formulae have been developed on the basis of an extensive reading comprehension experiment, aiming to evaluate the impact of a text simplification approach (a controlled language) on emergency instructions. Two evaluation scenarios have been presented, the first of which calculated with all three formulae, while the second used only the simplest one. In this way, the article aims to address both the lack of common TS evaluation metrics as suggested in Section 2 (Coster and Kauchak, 2011) and the scarcity of reading comprehension (Siddharthan and Katsos, 2012) evaluation with real users (Williams and Reiter, 2008), by proposing a tailored approach for this type of text simplification evaluation. With this article we aim at inciting the Text Simplification Community to open a discussion forum about common methods for evaluating text simplification, in order to provide objective evaluation metrics allowing the comparison of different approaches, and to ensure that simplification really achieves its aims. We also argue that taking in consideration the end-users and text units used for evaluation is important. In our approach, we address only the evaluation of text simplification approaches aiming to improve reading comprehension and experiments in which time to reply to questions and percentage of correct answers can be measured. A plausible scenario for applying our evaluation approach would be to use the Amazon Mechanical Turk for crowd-sourcing and then to evaluate the performance of a text simplification system on complex and simplified texts, to compare the performance of two or more approaches, or of two versions of the same system on the same pairs of texts. These formulae can be also employed in psycholinguistically-oriented experiments, which aim to reach cognitive findings regarding specific target reader groups, such as dyslexics or autis-

tic readers. Future work will involve the comparison of the above proposed evaluation metrics with any of the metrics already employed in the related work, such as the recent and classic readability formulae, eye-tracking, reading rate, human judges ratings, and others. We consider that content preservation and grammaticality are not necessary to be evaluated for this approach, as the simplified texts have been produced manually, by linguists, who were native speakers of English.

Acknowledgments

The authors would like to thank Prof. Dr. Petar Temnikov for the ideas and advices about the research methodology, Dr. Anke Buttner for the psycholinguistic counseling about the experiment design, including questions, answers and texts selection and the simplification method psychological validity, and Dr. Constantin Orasan and Dr. Le An Ha for the testing interface implementation. The research of Irina Temnikova reported in this paper was partially supported by the project AComIn "Advanced Computing for Innovation", grant 316087, funded by the FP7 Capacity Programme (Research Potential of Convergence Regions).

Finally, the authors would also like to thank the PITR 2013 reviewers for their useful feedback.

References

Jan Alexandersson, Peter Ljunglf, Kathleen F. McCoy, Brian Roark, and Annalu Waller, editors. 2012. *Proceedings of the Third Workshop on Speech and Language Processing for Assistive Technologies*. Association for Computational Linguistics, Montréal, Canada, June.

Stefan Bott, Luz Rello, Biljana Drndarević, and Horacio Saggion. 2012. Can spanish be simpler? lexsis: Lexical simplification for spanish. In *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012), Mumbai, India (December 2012)*.

Yvonne Canning. 2002. *Syntactic Simplification of Text*. Ph.D. thesis, University of Sunderland, UK.

Raman Chandrasekar, Christine Doran, and Bangalore Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 1041–1044. Association for Computational Linguistics.

James Clarke and Mirella Lapata. 2006. Models for sentence compression: A comparison across domains, training requirements and evaluation measures. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 377–384. Association for Computational Linguistics.

Jacob Cohen et al. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

William Coster and David Kauchak. 2011. Learning to simplify sentences using wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 1–9. Association for Computational Linguistics.

Walter Daelemans, Anja Höthker, and Erik Tjong Kim Sang. 2004. Automatic sentence simplification for subtitling in dutch and english. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1045–1048.

Jan De Belder and Marie-Francine Moens. 2012. A dataset for the evaluation of lexical simplification. In *Computational Linguistics and Intelligent Text Processing*, pages 426–437. Springer.

Siobhan Devlin. 1999. *Automatic Language Simplification for Aphasic Readers*. Ph.D. thesis, University of Sunderland, UK.

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

Biljana Drndarević, Sanja Štajner, Stefan Bott, Susana Bautista, and Horacio Saggion. 2013. Automatic text simplification in spanish: A comparative evaluation of complementing modules. In *Computational Linguistics and Intelligent Text Processing*, pages 488–500. Springer.

William H. DuBay. 2004. The principles of readability. *Impact Information*, pages 1–76.

Noémie Elhadad. 2006. *User-sensitive text summarization: Application to the medical domain*. Ph.D. thesis, Columbia University.

Rudolf Flesch. 1948. A new readability yardstick. *The Journal of applied psychology*, 32(3).

Caroline Gasperin, Erick Maziero, Lucia Specia, TAS Pardo, and Sandra M Aluisio. 2009. Natural language processing for social inclusion: a text simplification architecture for different literacy levels. *the Proceedings of SEMISH-XXXVI Seminário Integrado de Software e Hardware*, pages 387–401.

Gregory Grefenstette. 1998. Producing intelligent telegraphic text reduction to provide an audio scanning service for the blind. In *Working notes of the*

AAAI Spring Symposium on Intelligent Text summarization, pages 111–118.

Norman Edward Gronlund. 1982. *Constructing achievement tests*. Prentice Hall.

Kentaro Inui, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. 2003. Text simplification for reading assistance: a project note. In *Proceedings of the second international workshop on Paraphrasing-Volume 16*, pages 9–16. Association for Computational Linguistics.

Beata Beigman Klebanov, Kevin Knight, and Daniel Marcu. 2004. Text simplification for information-seeking applications. In *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE*, pages 735–747. Springer.

Irving Lorge. 1948. The lorge and flesch readability formulae: A correction. *School and Society*, 67:141–142.

Aurélien Max. 2006. Writing for language-impaired readers. In *Computational Linguistics and Intelligent Text Processing*, pages 567–570. Springer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Sarah E. Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *In Proc. of Workshop on Speech and Language Technology for Education*.

Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013. Simplify or help? text simplification strategies for people with dyslexia. *Proc. W4A*, 13.

J. Richard Ruffino. 1982. Coping with machine translation. *Practical Experience of Machine Translation*.

Advaith Siddharthan and Napoleon Katsos. 2012. Offline sentence processing measures for testing readability with users. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 17–24. Association for Computational Linguistics.

Advaith Siddharthan. 2003. *Syntactic simplification and text cohesion*. Ph.D. thesis, University of Cambridge, UK.

Matthew Snover, Nitin Madnani, Bonnie J Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or hter?: exploring different human judgments with a tunable mt metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268. Association for Computational Linguistics.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. Semeval-2012 task 1: English lexical simplification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 347–355. Association for Computational Linguistics.

Lucia Specia. 2010. Translating from complex to simplified sentences. In *Computational Processing of the Portuguese Language*, pages 30–39. Springer.

A. A. Streiff. 1985. New developments in titus 4. *Lawson (1985)*, 185:192.

Irina Temnikova. 2012. *Text Complexity and Text Simplification in the Crisis Management domain*. Ph.D. thesis, Wolverhampton, UK.

Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. 2007. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, 43(6):1606–1618.

David Vickrey and Daphne Koller. 2008. Sentence simplification for semantic role labeling. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-2008: HLT)*, pages 344–352.

Sandra Williams and Ehud Reiter. 2008. Generating basic skills reports for low-skilled readers. *Natural Language Engineering*, 14(4):495–525.

Sandra Williams, Advaith Siddharthan, and Ani Nenkova, editors. 2012. *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*. Association for Computational Linguistics, Montréal, Canada, June.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 409–420. Association for Computational Linguistics.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 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*, pages 365–368. Association for Computational Linguistics.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd international conference on computational linguistics*, pages 1353–1361. Association for Computational Linguistics.