Assessing Conformance of Manually Simplified Corpora with User Requirements: the Case of Autistic Readers

Sanja Štajner and **Richard Evans** and **Iustin Dornescu**
Research Group in Computational Linguistics
Research Institute of Information and Language Processing
University of Wolverhampton, UK
{SanjaStajner, R.J.Evans, I.Dornescu2}@wlv.ac.uk

Abstract

In the state of the art, there are scarce resources available to support development and evaluation of automatic text simplification (TS) systems for specific target populations. These comprise parallel corpora consisting of texts in their original form and in a form that is more accessible for different categories of target reader, including neurotypical second language learners and young readers. In this paper, we investigate the potential to exploit resources developed for such readers to support the development of a text simplification system for use by people with autistic spectrum disorders (ASD). We analysed four corpora in terms of nineteen linguistic features which pose obstacles to reading comprehension for people with ASD. The results indicate that the Britannica TS parallel corpus (aimed at young readers) and the Weekly Reader TS parallel corpus (aimed at second language learners) may be suitable for training a TS system to assist people with ASD. Two sets of classification experiments intended to discriminate between original and simplified texts according to the nineteen features lent further support for those findings.

1 Introduction

As a fundamental human right, people with reading and comprehension difficulties are entitled to access written information (UN, 2006). This entitlement enables better inclusion into society. However, the vast majority of texts that such people encounter in their everyday life – especially newswire texts – are lexically and syntactically very complex. Since the late nineties, several initiatives have emerged which propose guidelines for producing plain, easy-to-read and more accessible documents. These include the "Federal Plain Language Guidelines"[1], "Make it Simple, European Guidelines for the Production of Easy-to-Read Information for people with Learning Disability" (Freyhoff et al., 1998), "Am I making myself clear? Mencap's guidelines for accessible writing"[2], and the W3C – Web Accessibility Initiative guidelines[3]. However, manual adaptation of texts cannot match the speed with which new texts are published on the web in order to provide up to date information. The aim of Automatic Text Simplification (ATS) is to automatically (or at least semi-automatically) convert complex sentences into a more accessible form while preserving their original meaning. In the last twenty years, many ATS systems have been proposed for different target populations in various languages (Carroll et al., 1998; Devlin and Unthank, 2006; Saggion et al., 2011; Inui et al., 2003; Aluísio et al., 2008). Due to the scarcity of parallel corpora of original and manually simplified texts, most of these systems are rule-based.

The emergence of Simple English Wikipedia (SEW)[4], together with the existing English Wikipedia (EW)[5] provided a large amount of parallel TS training data, which motivated a shift in English TS from rule-based to data-driven approaches (Yatskar et al., 2010; Biran et al., 2011; Woodsend and Lapata, 2011; Coster and Kauchak, 2011; Wubben et al., 2012; Zhu et al., 2010). However, no assessment has

[1] http://www.plainlanguage.gov/howto/guidelines/bigdoc/fullbigdoc.pdf
[2] http://www.easy-read-online.co.uk/media/10609/making-myself-clear.pdf
[3] http://www.w3.org/WAI/
[4] http://simple.wikipedia.org/wiki/Main_Page
[5] http://wikipedia.org/wiki/Main_Page

Proceedings of the Workshop on Automatic Text Simplification: Methods and Applications in the Multilingual Society, pages 53–63,
Dublin, Ireland, August 24th 2014.

ever been made of the quality of the simplifications made in SEW and the usefulness of the transformations learned from EW–SEW parallel corpora for any of the specified target populations. The only instructions given to the authors of SEW are to use Basic English vocabulary and shorter sentences. The main page states that SEW is for everyone, including children and adults who are learning English. All previously mentioned studies conducted on that corpus evaluated the quality of the generated output in terms of grammaticality, meaning preservation, and simplicity, but not usefulness. Also, there have been no comparisons of the types of transformations present in EW–SEW with any of the other TS corpora in English which were simplified with a specific target population in mind, e.g. Encyclopedia Britannica and its manually simplified versions for children – Britannica Elementary (Barzilay and Elhadad, 2003)[6], Guardian Weekly and its manually simplified versions for language learners (Allen, 2009), and the FIRST corpus of various texts simplified for people with autism spectrum disorder (ASD)[7].

In this study, we compare the original and simplified texts of the four aforementioned TS corpora in terms of nineteen features which measure the complexity of texts for people with ASD. Although these features were derived from user requirements for people with ASD, many of them are known to present reading obstacles for other target populations as well (e.g. children or language learners). Given the lack of parallel TS corpora for people with ASD, our main goal is to investigate whether the EW–SEW or the other two corpora aimed at children and language learners could be used as training material for a TS system to assist people with ASD and thus enable data-driven approaches (instead of the currently used rule-based ones). In order to further support the results of this analysis, we conduct several classification experiments in which we try to distinguish between original and simplified texts in each of the four corpora, using the nineteen features.

2 The FIRST Project and User Requirements

Autistic Spectrum Disorders (ASD) are neurodevelopmental disorders characterised by qualitative impairment in communication and stereotyped repetitive behaviour. People with ASD show a diverse range of reading abilities: 5-10% have the capacity to read words from an early age without the need for formal learning (hyperlexia) but many demonstrate reduced comprehension of what has been read (Volkmar and Wiesner, 2009). They may have difficulty inferring contextual information or may have trouble understanding mental verbs, emotional language, and long sentences with complex syntactic structure (Tager-Flusberg, 1981; Kover et al., 2012). To address these difficulties, a tool is being developed in the FIRST project[8] to assist in the process of making texts more accessible for people with ASD. To achieve this, three modues are exploited:

1. **Structural complexity processor**, which detects syntactically complex sentences and generates alternatives to such sentences in the form of sequences of shorter sentences (Evans et al., 2014; Dornescu et al., 2013).

2. **Meaning disambiguator**, which resolves pronominal references, performs word sense disambiguation, and detects lexicalised (conventional) metaphors (Barbu et al., 2013).

3. **Personalised document generator**, which aggregates the output of processors 1 and 2 and generates additional elements such as glossaries, illustrative images, and document summaries.

The system, named *Open Book*, is deployed as an editing tool for healthcare and educational service providers. It functions semi-automatically, exploiting the three processors and requiring the user to authorise the application of the conversion operations. The system is required to assess the readability of texts, not only to decide which texts should be converted, but also to assess the readability of texts that are undergoing conversion. It is expected that people working to improve the accessibiity of a given text will benefit from relevant feedback concerning the effects of the changes being introduced. Automatic assessment of readability is one method by which such feedback can be delivered. In the

[6] http://www.cs.columbia.edu/ noemie/alignment/

[7] Available at: http://www.first-asd.eu/?q=system/files/FIRST_D7.2_20130228_annex.pdf

[8] www.first-asd.eu

context of improving the accessibility of texts, relevant feedback should indicate the extent to which different versions of a text meet the particular requirements of intended readers.

User requirements were obtained through consulatation of 94 subjects meeting the strict DSM-IV criteria for ASD and with IQ > 70. 43 user requirements were derived and assigned a reference code. The requirements link linguistic phenomena to editing operations, such as deletion, explanation, or transformation, that will convert the text to a more accessible form. The linguistic phenomena of concern include instances of syntactic complexity such as long sentences containing more than 15 words (possibly containing multiple copulative coordinated clauses (UR301), subordinate adjective clauses (UR302), explicative clauses (UR303), non-initial adverbial clauses (UR307)), sentences containing passive verbs (UR313), rarely used conjunctions and antithetic conjuncts (UR304, UR305, UR306), uncommon synonyms of polysemic words (UR401, UR425, UR504, UR505, UR511), rarely-used symbols and punctuation marks (UR311), anaphors, words containing more than 7 characters, adjectives ending with *-ly*, long numerical expressions (UR417), negation (UR314), words more than 7 characters long and adverbs with suffix *-ly* (UR317-319), anaphors, including pronouns (UR418-420).

Additional linguistic phenomena such as phraseological units (UR402, UR410, UR425, UR507), and non-lexicalised metaphors (UR422, UR508), were also found to pose obstacles to reading comprehension for people with ASD. At present, there is a scarcity of resources enabling accurate detection of these items. For this reason, changes in the prevalence of these items in original and converted versions of texts are not captured in this study. The full set of user requirements is detailed in Martos et al. (2013). More generally, it is infrequent linguistic phenomena that cause the greatest difficulty.

3 Related Work

There have been several studies analysing the existing TS corpora. However, their main focus was on determining necessary transformations in TS: for children (Bautista et al., 2011); for people with intellectual disability (Drndarević and Saggion, 2012); for language learners (Petersen and Ostendorf, 2007); and for people with low literacy (Gasperin et al., 2009). Unfortunately, those studies are not directly comparable (neither among themselves nor with our study), either because they focus on different types of transformations (the study of Bautista et al. (2011) focuses on general transformations while the other three studies focus on sentence transformations), or because they treat different languages (Spanish, English, and Brazilian Portuguese).

Two previous studies most relevant to ours are those by Napoles and Dredze (2010), and by Štajner et al. (2013). Napoles and Dredze (2010) built a statistical classification system that discriminates *simple* English from *ordinary* English, based on EW–SEW corpus. They used four different groups of features: lexical, part-of-speech, surface, and syntactic parse features. The accuracy of the best classifier (SVM) on the document classification task when using all features was 99.90%, while the accuracy of the best classifier (maximum entropy) on the sentence classification task when using all features was 80.80%. However, this study only demonstrated that it is fairly easy to discriminate sentences and documents of EW from those of SEW. It did not investigate whether the *simple* English used in SEW complies with the user requirements of any specific population with reading difficulties. Štajner et al. (2013) analysed a corpus of 37 newswire texts in Spanish and their manual simplifications aimed at people with Down's syndrome, compiled in the Simplext project[9]. They built a classification system that discriminates the original texts from those which are simple with an F-measure of 1.00 using the SVM, and only seven features: average number of punctuation marks (not counting end of sentence markers), numerical expressions, average word length in characters, the ratio of simple and complex sentences, sentence complexity index, lexical density and lexical richness. They reported the average sentence length as being the feature with the best discriminative power, leading to an F-measure of 0.99 when used on its own.

In spite of the many linguistic phenomena that pose obstacles to reading comprehension for different target populations, there have been almost no studies investigating whether a TS system built with a specific target population in mind could be successfully applied – or adapted – to a different target

[9] www.simplext.es

Corpus	Aimed at	Version	Code	Texts	SentPerText	WordsPerText
Weekly Reader	Language learners	Original	Learn.-O	100	39.41 ± 14.43	746.83 ± 174.25
		Simple	Learn.-S	100	38.40±12.59	621.11 ± 157.17
Enc. Britannica	Children	Original	Brit.-O	20	27.10 ± 8.91	628.30 ± 198.19
		Simple	Brit.-S	20	26.45 ± 9.35	382.35 ± 127.69
Wikipedia	Various	Original	Wiki-O	110	34.55 ± 1.87	716.57 ± 117.82
		Simple	Wiki-S	110	34.49 ± 1.82	675.07 ± 107.03
FIRST	People with ASD	Original	FIRST-O	25	13.64 ± 3.95	285.68 ± 34.46
		Simple	FIRST-S	25	22.92 ± 4.79	311.36 ± 76.82

Table 1: Corpora characteristics

population. The only exception to this is the study by Štajner and Saggion (2013), which demonstrated that two classifiers – one which discriminates sentences which should be split from those which should be left unsplit, and another which discriminates sentences which should be deleted from those which should be preserved – can successfully be trained on one type of corpora and applied to the other. Both corpora consisted of texts in Spanish, one containing newswire texts manually simplified for people with Down's syndrome, and the other various text genres manually simplified for people with ASD.

Motivated by those previous studies and the lack of parallel corpora aimed specifically to people with ASD, in this paper, we investigate whether some of already existing corpora for TS in English could potentially be used for building a data-driven TS system for this target population.

4 Methodology

The corpora, features, and experimental settings used in this study are described in Sections 4.1–4.3.

4.1 Corpora

Four parallel corpora of original and manually simplified texts for different target populations were used in this study (Table 1):

1. The corpus of 100 texts from *Weekly Reader* and their manual simplifications provided by Macmillan English Campus and Onestopenglish[10] aimed at foreign language learners. The corpus is divided into three sub-corpora – advanced, intermediate and elementary – each representing a different level of simplification. Given that the other three corpora used in this study contain original texts and only one level of simplification, we only used the texts from the advanced (henceforth *original*) and elementary (henceforth *simplified*) levels. A more detailed description of this corpus can be found in (Allen, 2009).

2. The corpus of 20 texts from the Encyclopedia Britannica and their manually simplified versions aimed at children – Britannica Elementary (Barzilay and Elhadad, 2003)[11].

3. The corpus of 110 randomly selected corresponding articles from EW and SEW. Here, it is important to note that, in general, articles from SEW do not represent direct simplifications of the articles from EW, they just have a matching topic. For this reason, we did not use complete EW and SEW articles. We only used those sentences in original and simplified versions, which existed in the sentence-aligned parallel corpora version 2.0[12] (Kauchak, 2013).

4. The corpus of 25 texts on various topics manually simplified for people with autism, compiled in the FIRST project[13], for the purpose of a piloting task[14]. The texts were simplified by carers of people with ASD in accordance with specified guidelines.

[10] http://www.onestopenglish.com/
[11] http://www.cs.columbia.edu/ noemie/alignment/
[12] http://www.cs.middlebury.edu/ dkauchak/simplification/
[13] www.first-asd.eu
[14] http://www.first-asd.eu/?q=system/files/FIRST_D7.2_20130228_annex.pdf

4.2 Text Features Relevant to User Requirements

In this paper, a set of 15 text complexity measures and 4 formulae exploiting these measures was used to estimate the accessibility of the texts. These features quantify the occurrence of linguistic phenomena identified as potential obstacles to reading comprehension for people with ASD. The set of features is presented in Table 2. The set of formulae is presented in Table 3. In every case, accessible texts are expected to have smaller values of each metric.

#	Code	Linguistic feature	Explanation/relevance
1	Illative	Illative conjunctions	Indicators of syntactic complexity, linking clauses.
2	CompConj	Comparative conjunctions	[UR304-306]
3	AdvConj	Adversative conjunctions	
4	LongSent	Long sentences	Motivated by the assumption that deriving the propositions in
5	Semicol	Semicolons/suspension points	complex sentences is more difficult than deriving connections between related propositions expressed in simple sentences
6	Passive	Passive verbs	(Arya et al., 2011). [UR309-310, UR313]
7	UnPunc	Unusual punctuation	Indicates syntactic complexity, ellipsis, alternatives, and mathematical expressions [UR311]
8	Negations	Negation	The sum of adverbial and morphological negations ("Make it Simple" (Freyhoff et al., 1998), though contrary to the findings of Tattamanti (2008)) [UR314]
9	Senses	Possible senses	The sum over all tokens in the text of the total number of possible senses of each token. [UR401, UR425, UR504-505, UR511]
10	PolyW	Polysemic words	Words with two or more senses listed in WordNet. [UR401, UR425, UR504, UR505, UR511]
11	Infreq	Infrequent words	Words that are not among the 5000 most frequent words in English [UR304-306, UR401, UR425, UR504-505, UR511]
12	NumExp	Numerical expressions	Numbers written as sequences of words rather than digits [UR417]
13	Pron	Pronouns	Studies have shown that people with ASD can have
14	DefDescr	Definite descriptions	difficulty processing anaphora (Fine et al., 1994) [UR418-420]
15	SylLongW	Long words	Words with more than three syllables [UR317-319]

Table 2: Complexity measures (1 – words such as *therefore* and *hence*; 2 – words such as *equally* and *correspondingly*; 3 – words such as *although* and *conversely*; 4 – sentences more than 15 words long; 8 – negative adverbials and negative prefixes such as *un-* and *dis-*; 11 – derived from Wiktionary frequency lists for English[16])

#	Code	Metric	Formula	Relevance
16	PolyType	Polysemic type ratio	$\frac{ptyp}{typ}$	Indicates the proportion of the text vocabulary that is polysemous. [UR401, UR425, UR504-505, UR511]
17	CommaInd	Comma index	$\frac{10 \times c}{w}$	Indicates the average syntactic complexity of the sentences in the text [UR301-303, UR307]
18	WordsPerSent	Words per sentence	$\frac{w}{s}$	Indicates the average length of the sentences in the text [UR309]
19	TypeTokRat	Type-token ratio	$\frac{typ}{tok}$	Indicate the range of vocabulary used in the text [UR401, UR425, UR504, UR505, UR511]

Table 3: Text complexity formulae (w – the number of words in the text; s – the number of sentences in the text; $ptyp$ – the number of polysemic word types in the text; c – the number of commas in the text; typ – the number of word types in the text; tok – the number of word tokens in the text)

Scores for these measures, and the text complexity formulae that exploit them where obtained automatically by the tokeniser, part-of-speech tagger, and lemmatiser distributed with LT TTT2 (Grover et al., 2000). Detection of the features used to derive complexity measures also involved the use of additional resources such as WordNet, gazetteers of rare illative, comparative, and adversative conjunctions, negatives (words and prefixes) and a set of lexico-syntactic patterns used to detect passive verbs (presented in Figure 1).

am/are/is/was/were w_{RB}* $w_{\{VBN|VBD\}}$
am/are/is/was/were w_{RB}* *being* w_{RB}* $w_{\{VBN|VBD\}}$
have/has/had w_{RB}* *been* w_{RB}* $w_{\{VBN|VBD\}}$
will w_{RB}* *be* w_{RB}* $w_{\{VBN|VBD\}}$
am/is/are w_{RB}* *going* w_{RB}* *to* w_{RB}* *be* w_{RB}* $w_{\{VBN|VBD\}}$
w_{MD} w_{RB}* *be* $w_{\{VBN|VBD\}}$
w_{MD} w_{RB}* *have* w_{RB}* *been* w_{RB}* $w_{\{VBN|VBD\}}$

Figure 1: Lexico-syntactic patterns used to detect passive verbs ('*' indicates zero or more repetitions of the item it is attached to, while RB, VBN, VBD, and MD are Penn treebank tags returned by the LT TTT PoS tagger: RB – adverb; VBN – past participle; VBD – past tense; and MD – modal verb)

4.3 Experiments

Two sets of experiments were performed in this study:

1. Analysis of differences between original and simplified texts in terms of nineteen selected features (Section 4.2) across four corpora (Section 4.1). Statistical difference was measured using the t-test for related samples in the cases where the features were normally distributed, and using the related samples Wilcoxon signed rank test otherwise. Normality of the data was tested using the Shapiro-Wilk test of normality, which is preferred over the Kolmogorov-Smirnov test when the dataset contains less than 2,000 elements. All tests were performed in SPSS. Features 1–15 were first normalised (as an average per sentence) in order to allow a fair comparison across the four TS corpora (text length in words and sentences differed significantly across different corpora).

2. Classification experiments with the aim of discriminating original from simplified texts using the nineteen selected features. All experiments were conducted using the Weka Experimenter (Witten and Frank, 2005; Hall et al., 2009) in 10-fold cross-validation setup with 10 repetitions, using four different classification algorithms: NB – NaiveBayes (John and Langley, 1995), SMO – Weka implementation of Support Vector Machines (Keerthi et al., 2001) with normalisation, JRip – a propositional rule learner (Cohen, 1995), and J48 – Weka implementation of C4.5 (Quinlan, 1993). The statistical significance of the observed differences in F-measures obtained by different algorithms was calculated using the corrected paired t-test provided in the Weka Experimenter.

The TS system in FIRST is not only supposed to decide which texts should be converted, but also to assess the readability of texts that are undergoing conversion. It is expected that people working to improve the accessibility of a given text will benefit from relevant feedback concerning the effects of the changes being introduced. Automatic assessment of readability is one method by which such feedback can be delivered. Deriving a subset of features which, when trained with an appropriate classification algorithm, can categorize a given text as either 'original' or 'simplified', would facilitate automatic evaluation of TS systems. The resulting classifier would be suitable for assessing whether those systems perform an appropriate level of simplification. This could serve as a rough estimation, an efficient first step offering a quick evaluation prior to being tested with real users.

5 Results and Discussion

The results of the two sets of experiments are presented and discussed in the next two subsections.

5.1 Analysis of the Features across the Corpora

Mean values (with standard deviations) of each of the first eight features on each sub-corpus are displayed in Table 4. The number of unusual punctuation marks (*UnPunc*) is the only feature whose value does not differ significantly between the original and simplified versions of the texts in any of the four corpora. This feature was thus excluded from further classification experiments. The number of comparative conjunctions per sentence (*CompConj*) significantly decreases only when simplifying texts for

Corpus	Illative	CompConj	AdvConj	LongSent	Semicol	UnPunc	Passive	Negations
Lear.-O	0.24±0.12	0.04±0.13	0.21±0.08	0.62±0.15	0.03±0.05	0.00±0.01	0.21±0.10	0.33±0.15
Lear.-S	**0.20±0.13**	0.03±0.09	**0.19±0.09**	**0.51±0.14**	***0.03±0.05**	0.00±0.01	**0.09±0.09**	**0.26±0.14**
Brit.-O	0.13±0.09	0.15±0.26	0.14±0.07	0.72±0.11	0.13±0.20	0±0	0.33±0.10	0.28±0.16
Brit.-S	**0.08±0.05**	***0.02±0.10**	**0.06±0.04**	**0.38±0.11**	**0.00±0.02**	0±0	**0.25±0.12**	**0.14±0.09**
Wiki-O	0.20±0.11	0.11±0.19	0.16±0.10	0.65±0.12	0.04±0.04	0.04±0.10	0.34±0.15	0.32±0.23
Wiki-S	**0.18±0.11**	0.11±0.20	**0.14±0.09**	**0.62±0.12**	**0.03±0.04**	0.03±0.10	0.33±0.15	**0.29±0.24**
FIRST-O	0.18±0.14	0.06±0.19	0.18±0.15	0.68±0.15	0.03±0.10	0.01±0.02	0.27±0.23	0.42±0.28
FIRST-S	**0.11±0.10**	0.01±0.06	**0.09±0.07**	**0.33±0.19**	0.00±0.01	0±0	0.20±0.15	**0.22±0.13**

Table 4: Mean values (with standard deviation) of features 1–8 across the corpora (*O* – the original texts in the corpora; *S* – the simplified texts in the corpora; **bold** – significantly different from the value on the original texts at a 0.01 level of significance; ***bold** – significantly different from the value on the original texts at a 0.05 level of significance (but not at 0.01); '0.00' – a value different from zero which rounded at two decimals gives 0.00; '0' – a value equal to zero)

Corpus	Senses	PolyW	Infreq	NumExp	Pron	DefDescr	SylLongW
Lear.-O	73.95±12.32	9.37±1.72	5.64±1.33	0.18±0.11	0.97±0.40	1.86±0.54	1.12±0.28
Lear.-S	**64.21±11.16**	**7.85±1.45**	**4.14±1.01**	**0.16±0.10**	**0.90±0.37**	**1.62±0.45**	**0.92±0.27**
Brit.-O	67.51± 8.83	9.87±1.15	9.37±1.10	0.18±0.12	0.40±0.18	2.86±0.44	1.45±0.20
Brit.-S	**48.68± 4.17**	**6.48±0.57**	**5.39±0.58**	**0.09±0.06**	**0.28±0.13**	**1.86±0.20**	**1.17±0.19**
Wiki-O	67.70±12.96	9.13±1.61	7.86±1.63	0.18±0.16	0.67±0.43	2.08±0.58	1.24±0.38
Wiki-S	68.20±13.56	**8.71±1.56**	**7.16±1.51**	***0.17±0.16**	0.68±0.44	**1.97±0.54**	**1.10±0.42**
FIRST-O	82.28±24.20	10.16±2.65	7.11±2.72	0.19±0.19	1.05±0.73	2.12±0.92	1.17±0.58
FIRST-S	**57.13±15.96**	**6.47±1.77**	**3.92±1.56**	**0.09±0.07**	***0.82±0.44**	**1.62±0.54**	***0.92±0.43**

Table 5: Mean values (with standard deviation) of features 9–15 across the corpora (*O* – the original texts in the corpora; *S* – the simplified texts in the corpora; **bold** – significantly different from the value on the original texts at a 0.01 level of significance; ***bold** – significantly different from the value on the original texts at a 0.05 level of significance (but not at 0.01))

children (*Brit.-S*), while the average number of passive constructions per sentence (*Passive*) decreases when simplifying for both children (*Brit.-S*) and language learners (*Lear.-S*). It is interesting to note that the average number of passive constructions per sentence (*Passive*) does not decrease in the EW–SEW corpus and that its value on the simplified versions of Wikipedia articles (*Wiki-S*) is significantly higher than on *Brit.-S* and *Lear.-S*, although SEW claims to provide articles simplified for both those target populations. It can also be observed that the fact that all four corpora were reported to have significant differences between original and simplified texts in terms of features *Illative*, *AdvConj*, *LongSent*, and *Negations* does not necessarily mean that the average number of occurrences of those features is similar in all four simplified corpora. The values of *Illative*, *AdvConj*, and *LongSent* in the simplified versions of the texts in the FIRST corpus seem to correspond best to those in the simplified versions of the texts in the Britannica corpus (*Brit.-S*). The value of *Negations* in *FIRST-S*, however, seems to correspond best to that in *Lear.-S*. This suggests that if we wish to build a component of our TS system (to assist people with ASD) which would remove negations (*Negations*), we should train it on the sentence pairs from the corpora with simplifications aimed at second language learners. If we wish to build a component which would remove illative conjunctions (*Illative*), adversative conjuctions (*AdvConj*), or long sentences (*LongSent*), we should probably train it on the sentence pairs from the corpora with simplifications aimed at young readers.

The number of occurrences per sentence of features 9–15 in the original versions of the texts was significantly higher than in the simplified versions of the texts in all four corpora, with only two exceptions – features *Senses* and *Pron* in the EW–SEW corpus (*Wiki-O* and *Wiki-S*), as can be observed in Table 5. Again, the mean values of all features in the simplified versions of the texts in the FIRST corpora *FIRST-S*, seems to correspond better to the simplified versions of Encyclopedia Britannica (*Brit.-S*) and

Corpus	PolyType	CommaInd	WordsPerSent	TypeTokRat
Lear.-O	0.76±0.04	0.56±0.12	19.91±3.46	0.51±0.04
Lear.-S	**0.77±0.04**	**0.46±0.15**	**16.69±2.78**	**0.47±0.05**
Brit.-O	0.69±0.03	0.78±0.15	23.46±2.78	0.51±0.04
Brit.-S	***0.71±0.02**	***0.67±0.14**	**14.61±1.21**	**0.55±0.04**
Wiki-O	0.71±0.05	0.65±0.15	20.73±3.16	0.48±0.05
Wiki-S	**0.71±0.05**	**0.60±0.16**	**19.57±2.90**	***0.48±0.05**
FIRST-O	0.73±0.04	0.51±0.18	22.20±5.43	0.59±0.05
FIRST-S	0.75±0.06	**0.19±0.15**	**13.86±3.41**	**0.53±0.08**

Table 6: Mean values (with standard deviation) of features 16–19 across the corpora (*O* – the original texts in the corpora; *S* – the simplified texts in the corpora; **bold** and ***bold** – used in the same way as in the previous two tables)

Weekly Readers (*Lear.-S*) than to those in the simplified versions of the Wikipedia articles (*Wiki-S*). It is also interesting to note that many of the features (*LongSent*, *Negations*, *Senses*, *PolyW*, *Infreq*, *DefDesc*) seem to have a significantly higher number of occurrences per sentence in the simplified versions of the Wikipedia articles (*Wiki-S*) than in the simplified versions of Encyclopedia Britannica (*Brit.-S*) and Weekly Reader (*Lear.-S*).

The comma index (*CommaInd*), type-token ratio (*TypeTokRat*), and the average number of words per sentence (*WordsPerSent*) were found to be significantly higher in original texts than in their simplified versions in all four corpora (Table 6). However, the values of those three text complexity formulae were not similar in the simplified texts across the four corpora. In terms of the average number of words per sentence (*WordsPerSent*) and the type-token ratio (*TypeTokRat*), the simplified versions of the texts in the FIRST corpora (*FIRST-S*) seem to correspond better to the texts simplified for young readers (*Brit.-S*), than to those simplified for second language learners (*Lear.-S*) and those aimed at various target populations (*Brit.-S*). The comma index (*CommaInd*) obtained for simplified texts in the FIRST corpora was several times lower than that obtained for simplified texts in the three other corpora. The polysemic type ratio (*PolyType*) was not significantly different in original and in simplified texts of the FIRST corpora (Table 6). The higher polysemic type ratio (*PolyType*) for simplified rather than original versions of the texts in the other three corpora was unexpected, as it is usually assumed that polysemous words can pose an obstacle for various target populations. However, it is important to bear in mind that polysemous words usually pose an obstacle when conveying one of their infrequently used meanings. Findings in cognitive psychology indicate that the words with the highest number of possible meanings are actually understood more quickly, due to their high frequency (Jastrzembski, 1981). A common lexical simplification strategy is to replace infrequent words with their more frequent synonyms, and long words with their shorter synonyms. This strategy leads to a higher polysemic type ratio (*PolyType*) in simplified versions of the texts as the shorter words are usually more frequent (Balota et al., 2004), and frequent words tend to be more polysemous than infrequent ones (Glanzer and Bowles, 1976).

5.2 Classification between Original and Simplified Texts

Classification experiments were conducted using two different sets of features on each of the corpora:

1. *all* – all 18 features (UnPunc was excluded as it was not reported as significant for any of the corpora)

2. *best* – 11 features which were reported as significant for all four corpora (Illative, AdvConj, LongSent, Negations, PolyW, NumExp, DefDescr, SylLongW, CommaInd, WordsPerSent, TypeTokRat)

As can be observed from Table 7, use of the SMO-n classification algorithm using the subset of 11 *best* features achieves perfect 1.00 F-measure for discriminating original from simplified versions of the Encyclopedia Britannica. The same classification algorithm performs less well on the FIRST and Weekly Readers corpora (though still quite well), while it performs significantly worse on the Wikipedia corpus.

The baseline (which chooses majority class) would be 0.50 in all cases. These results indicate that the Encyclopedia Britannica TS parallel corpus, and possibly the Weekly Readers TS parallel corpus, may serve as suitable training material for building a TS system (or at least some of its components) aimed at people with ASD.

Dataset	SMO-n	NB	JRip	J48
Brit-all	0.98±0.09	0.94±0.12	0.94±0.14	0.97±0.11
Brit-best	1.00±0.00	0.99±0.05	0.94±0.13	0.97±0.11
FIRST-all	0.88±0.15	0.86±0.19	0.79±0.23	0.75±0.25
FIRST-best	0.88±0.15	0.85±0.20	0.78±0.25	0.76±0.25
Lear-all	0.81±0.08	0.74±0.10*	0.75±0.07*	0.72±0.10*
Lear-best	0.77±0.08	0.74±0.11	0.70±0.10*	0.73±0.10
Wiki-all	0.54±0.12	0.50±0.12	0.51±0.14	0.35±0.20*
Wiki-best	0.55±0.13	0.55±0.12	0.51±0.12	0.33±0.20*

Table 7: F-measure with standard deviation in a 10-fold cross-validation setup with 10 repetitions for four classification algorithms: SMO-n, NB, JRip, and J48 (* – statistically significant degradation in comparison with SMO-n)

6 Conclusions

Automatic Text Simplification (ATS) aims to convert complex texts into a simpler form, which is more accessible to a wider audience. Due to the lack of parallel corpora for TS consisting of original and manually simplified texts, most of the ATS systems for specific target populations are still rule-based. Our main goal was to explore whether some of the existing TS parallel corpora in English, aimed at different audiences (children – Encyclopedia Britannica, language learners – Weekly Reader, and various – Wikipedia) could be used as training material to build a TS system aimed at people with ASD. We analysed the four corpora (FIRST, Britannica, Weekly Reader, and Wikipedia) in terms of nineteen linguistic features which pose obstacles to reading comprehension for people with ASD. The preliminary results indicate that the Britannica TS parallel corpus, and possibly the Weekly Reader TS parallel corpus, could be used to train a TS system aimed at people with ASD. Two sets of classification experiments which tried to discriminate original from simplified texts according to the nineteen features derived from user requirements further supported those findings. The results of the classification experiments indicated that the SVM classifier trained on the Britannica corpus might be suitable for discriminating original from simplified texts for people with ASD, and thus might be used as the initial evaluation of the texts simplified by the TS system developed in the FIRST project.

Acknowledgements

The research described in this paper was partially funded by the European Commission under the Seventh (FP7-2007-2013) Framework Programme for Research and Technological Development (FP7-ICT-2011.5.5 FIRST 287607).

References

D. Allen. 2009. A Corpus-Based Study of the Role of Relative Clauses in the Simplification of News Texts for Learners of English. *System*, 37(4):585–599.

S. M. Aluísio, L. Specia, T. A.S. Pardo, E. G. Maziero, and R. P.M. Fortes. 2008. Towards brazilian portuguese automatic text simplification systems. In *Proceedings of the eighth ACM symposium on Document engineering*, DocEng '08, pages 240–248, New York, NY, USA. ACM.

D. J. Arya, Elfrieda H. Hiebert, and P. D. Pearson. 2011. The effects of syntactic and lexical complexity on the comprehension of elementary science texts. *International Electronic Journal of Elementary Education*, 4 (1):107–125.

D. Balota, M. J. Cortese, S. D. Sergent-Marshall, D. H. Spieler, and M. J. Yap. 2004. Visual word recognition of single-syllabe words. *Journal of Experimental Psychology: General*, 133:283–316.

E. Barbu, M. Martín-Valdivia, L. Alfonso, and U. Lopez. 2013. Open book: a tool for helping asd users' semantic comprehension. In *Proceedings of the 2th Workshop of Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, pages 11–19, Atlanta, US. Association for Computational Linguistics.

R. Barzilay and N. Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, EMNLP '03, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

S. Bautista, C. León, R. Hervás, and P. Gervás. 2011. Empirical identification of text simplification strategies for reading-impaired people. In *European Conference for the Advancement of Assistive Technology*.

O. Biran, S. Brody, and N. Elhadad. 2011. Putting it Simply: a Context-Aware Approach to Lexical Simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 496–501, Portland, Oregon, USA. Association for Computational Linguistics.

J. Carroll, G. Minnen, Y. Canning, S. Devlin, and J. Tait. 1998. Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *Proceedings of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10.

W. Cohen. 1995. Fast Effective Rule Induction. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123.

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

S. Devlin and G. Unthank. 2006. Helping aphasic people process online information. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, Assets '06, pages 225–226, New York, NY, USA. ACM.

I. Dornescu, R. Evans, and C. Orasan. 2013. A Tagging Approach to Identify Complex Constituents for Text Simplification. In *Proceedings of Recent Advances in Natural Language Processing*, pages 221 – 229, Hissar, Bulgaria.

B Drndarević and H. Saggion. 2012. Reducing Text Complexity through Automatic Lexical Simplification: an Empirical Study for Spanish. *SEPLN Journal*, 49.

R. Evans, C. Orasan, and I. Dornescu. 2014. An evaluation of syntactic simplification rules for people with autism. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 131–140, Gothenburg, Sweden, April. Association for Computational Linguistics.

J. Fine, G. Bartolucci, P. Szatmari, and G. Ginsberg. 1994. Cohesive discourse in pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, 24:315–329.

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

C. Gasperin, L. Specia, T. Pereira, and S.M. Aluísio. 2009. Learning When to Simplify Sentences for Natural Text Simplification. In *Proceedings of the Encontro Nacional de Inteligncia Artificial (ENIA-2009), Bento Gonalves, Brazil.*, pages 809–818.

M. Glanzer and N. Bowles. 1976. Analysis of the word frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory*, 2:21–31.

C. Grover, C. Matheson, A. Mikheev, and M. Moens. 2000. Lt ttt - a flexible tokenisation tool. In *In Proceedings of Second International Conference on Language Resources and Evaluation*, pages 1147–1154.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18, November.

K. Inui, A. Fujita, T. Takahashi, R. Iida, and T. Iwakura. 2003. Text simplification for reading assistance: a project note. In *Proceedings of the second international workshop on Paraphrasing - Volume 16*, PARAPHRASE '03, pages 9–16, Stroudsburg, PA, USA. Association for Computational Linguistics.

J. Jastrzembski. 1981. Multiple meaning, number or related meanings, frequency of occurrence and the lexicon. *Cognitive Psychology*, 13:278–305.

G. H. John and P. Langley. 1995. Estimating Continuous Distributions in Bayesian Classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345.

D. Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1537–1546, Sofia, Bulgaria, August. Association for Computational Linguistics.

S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy. 2001. Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation*, 13(3):637–649.

S. T. Kover, E. Haebig, A. Oakes, A. McDuffie, R. J. Hagerman, and L. Abbeduto. 2012. Syntactic comprehension in boys with autism spectrum disorders: Evidence from specific constructions. In *Proceedings of the 2012 International Meeting for Autism Research*, Athens, Greece. International Society for Autism Research.

J. Martos, S. Freire, A. González, D. Gil, R. Evans, V. Jordanova, A. Cerga, A. Shishkova, and C. Orasan. 2013. FIRST Deliverable - User preferences: Updated. Technical Report D2.2, Deletrea, Madrid, Spain.

C. Napoles and M. Dredze. 2010. Learning simple wikipedia: a cogitation in ascertaining abecedarian language. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids*, CL&W '10, pages 42–50, Stroudsburg, PA, USA. Association for Computational Linguistics.

S. E. Petersen and M. Ostendorf. 2007. Text Simplification for Language Learners: A Corpus Analysis. In *Proceedings of Workshop on Speech and Language Technology for Education*.

R. Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

H. Saggion, E. Gómez Martínez, E. Etayo, A. Anula, and L. Bourg. 2011. Text Simplification in Simplext: Making Text More Accessible. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 47:341–342.

H. Tager-Flusberg. 1981. Sentence comprehension in autistic children. *Applied Psycholinguistics*, 2:1:5–24.

M. Tattamanti, R. Manenti, P. A. Della Rosa, A. Falini, D. Perani, S. Cappa, and A. Moro. 2008. Negation in the brain: Modulating action representations. *NeuroImage*, 43 (2008):358–367.

UN. 2006. Convention on the rigths of persons with disabilities.

F. R. Volkmar and L. Wiesner. 2009. *A Practical Guide to Autism*. Wiley, Hoboken, NJ, 2nd edition.

S. Štajner and H. Saggion. 2013. Adapting Text Simplification Decisions to Different Text Genres and Target Users. *Procesamiento del Lenguaje Natural*, 51:135–142.

S. Štajner, B. Drndarević, and H. Saggion. 2013. Corpus-based Sentence Deletion and Split Decisions for Spanish Text Simplification. *Computación y Systemas*, 17(2):251–262.

I. H. Witten and E. Frank. 2005. *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann Publishers.

K. Woodsend and M. Lapata. 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)*.

S. Wubben, A. van den Bosch, and E. Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 1015–1024, Stroudsburg, PA, USA. Association for Computational Linguistics.

M. Yatskar, B. Pang, C. Danescu-Niculescu-Mizil, and L. 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*, HLT '10, pages 365–368, Stroudsburg, PA, USA. Association for Computational Linguistics.

Z. Zhu, D. Berndard, and I. Gurevych. 2010. A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361.