Combining Multiple Corpora for Readability Assessment for People with Cognitive Disabilities

Victoria Yaneva, Constantin Orăsan, Richard Evans, and Omid Rohanian
Research Institute in Information and Language Processing,
University of Wolverhampton, UK
{v.yaneva, c.orasan, r.j.evans, omid.rohanian}@wlv.ac.uk

Abstract

Given the lack of large user-evaluated corpora in disability-related NLP research (e.g. text simplification or readability assessment for people with cognitive disabilities), the question of choosing suitable training data for NLP models is not straightforward. The use of large generic corpora may be problematic because such data may not reflect the needs of the target population. At the same time, the available user-evaluated corpora are not large enough to be used as training data. In this paper we explore a third approach, in which a large generic corpus is combined with a smaller population-specific corpus to train a classifier which is evaluated using two sets of unseen user-evaluated data. One of these sets, the ASD Comprehension corpus, is developed for the purposes of this study and made freely available. We explore the effects of the size and type of the training data used on the performance of the classifiers, and the effects of the type of the unseen test datasets on the classification performance.

1 Introduction

When developing educational tools and applications for students with cognitive disabilities, it is necessary to match the readability of the educational materials to the abilities of the students and to adapt the text content to their needs. Both text adaptation and readability research for people with cognitive disabilities are thus dependent on evaluation involving target users. However, there are two main difficulties in collecting data from users with cognitive disabilities: i) experiments involving those users are expensive to perform and ii) the task of text evaluation is challenging for target users because of their cognitive disability.

Following from the first difficulty, user-evaluated data is scarce and the majority of NLP research for disabled groups is done by exploiting ratings or simplification provided by teachers and experts (Inui et al., 2001; Dell'Orletta et al., 2011; Jordanova et al., 2013). Examples of such a corpora are the FIRST corpus (Jordanova et al., 2013), which contains 31 original articles and versions of the articles that had been manually simplified for people with autism, and a corpus of manually simplified sentences for congenitally deaf Japanese readers (Inui et al., 2001). Henceforth in this paper, we refer to such manually simplified corpora as *population-specific corpora*. These corpora have not been evaluated by end users with disabilities.

As a result of the second difficulty, the fact that people with cognitive disabilities find text evaluation challenging, the size of user-evaluated datasets is rather limited. For example, to the best of our knowledge, there is currently only one readability corpus evaluated by people with intellectual disability, called LocalNews (Feng, 2009). This corpus contains 11 original and 11 simplified news stories. In this paper we present another corpus evaluated by people with autism containing a total of 27 documents. Henceforth in the paper, we refer to these type of corpora as *user-evaluated corpora*.

Given the lack of large population-specific or user-evaluated corpora in disability-related research, the question of choosing suitable training data for NLP models is not straightforward. While the use of large generic corpora as training data may be inadequate as such data may not reflect the needs of the target population, the use of population-specific and user-evaluated corpora as training data is problematic due to the scarcity

of such data. In this paper we explore a third approach, in which a large generic corpus is combined with a smaller population-specific corpus to train

a system to predict the difficulty of text for people with autism. We compare the performance of this approach to: i) an approach exploiting only the large generic corpus and ii) an approach exploiting only the small population-specific corpus. We also compare the performance of the classification models derived from two different machine learning algorithms. All classifiers trained on the different corpora are then evaluated on two small sets of user-evaluated corpora (unseen data), one of which was developed for the purpose of this study (Section 3).

Contributions We developed the ASD Comprehension Corpus containing 27 educational articles evaluated by readers with autism and classified as *easy* and *difficult* based on participants' answers to comprehension questions. The texts and the answers of each participant for each question are currently available at: https://github.com/victoria-ianeva/ASD-Comprehension-Corpus[1]. Further, we explore i) the effects of the size and type of the training data on the external validity of the classifiers and ii) the effects of the type of unseen test datasets (only original versus original + simplified articles) on the classification performance. The system used in these experiments is available at: http://rgcl.wlv.ac.uk/demos/autor_readability

The rest of this paper is organised as follows. The next section presents related work relevant to this research, while Section 3 describes the process for the development of the ASD Comprehension corpus. Section 4 describes the corpora used in the study. Section 5 presents the derivation of the classification models, and Section 6 presents a discussion of the main findings, which are summarised in Section 7.

[1]The repository also contains the answers of participants from a control group (without autism), which were not explored in this article but may be useful to the community for investigating between-group differences. For more information about the control group see Yaneva (2016).

2 Related Work

Previous work from the fields of psycholinguistics, pertaining to language and autism, readability assessment, and domain adaptation are relevant to the research presented in our current paper.

2.1 Autism Spectrum Disorder

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition affecting communication and social interaction. The reading difficulties of some people with ASD include, but are not limited to, difficulties resolving ambiguity in meaning (Happé and Frith, 2006; Happe, 1997; Frith and Snowling, 1983; O'Connor and Klein, 2004), difficulties comprehending abstract words (Happé, 1995), difficulties in the syntactic processing of long sentences (Whyte et al., 2014), difficulties identifying the referents of pronouns (O'Connor and Klein, 2004), difficulties in figurative language comprehension (MacKay and Shaw, 2004), and difficulties in making pragmatic inferences (Norbury, 2014). Adults with autism have also been shown to process images inserted in easy-to-read documents differently from non-autistic control participants (Yaneva et al., 2015).

2.2 Readability Assessment

Readability is a construct which has been defined as the ease of comprehension because of the style of writing (Harris and Hodges, 1995). Historically, the readability of texts has been estimated via formulae exploiting shallow features such as word and sentence length (Dubay, 2004); cognitive models exploiting features such as age of acquisition of words and text cohesion (McNamara et al., 2014) and, more recently, thanks to advances in Natural Language Processing (NLP), readability has also been estimated via computational models (Collins-Thompson, 2014; François, 2015). Advances in the fields of NLP and Artificial Intelligence have enabled both the faster computation of existing statistical features and the development of new NLP-enhanced features (e.g., average parse-tree height, average distance between pronouns and their anaphors, etc.) which can be used in more complex methods of assessment based on machine learning. An example of a readability model targeted to a specific application of readability assessment are the unigram models by Si and Callan (2001), which have been found particularly suitable for assessment of Web con-

tent, where the presence of links, email addresses and other elements biases the traditional formulae.

In terms of readability assessment for readers with cognitive disabilities, previous research has shown that readability features such as *entity density per sentence* and *lexical chains* (synonymy or hyponymy relations between nouns) are useful for estimating the readability of texts for readers with mild intellectual disability (Feng et al., 2010). This is due to the fact that these readers struggle to remember relations that hold within- and between-sentences (Feng et al., 2010). Similarly, features such as *word length* or *word frequency* are more relevant for readability assessment for people with dyslexia because they struggle with decoding particular letter and syllable combinations (Rello et al., 2012). In the case of autism, an important issue has been the lack of corpora whose reading difficulty levels have been evaluated by people with autism. For this reason most readability research for this population has so far focused on texts simplified by experts (Štajner et al., 2014). User-evaluated texts were used for the first time in a study, where the discriminatory power of a number of features was evaluated on a preliminary dataset of 16 texts considered easy or difficult to comprehend by people with autism (Yaneva and Evans, 2015).

2.3 Domain adaptation

Supervised machine learning and statistical methods like the ones used in this paper benefit from the availability of large amounts of training data. However, in many cases it is not easy to obtain enough training data for specific domains or applications. As a result it is not uncommon that researchers train on data from one domain and test on data from a different one. As would be expected, this usually leads to lower levels of performance. The field of domain adaptation is addressing this problem by proposing methods that can perform well even when the training and testing domains are different. In many cases this is achieved by exploiting a small training corpus of the same domain as the test documents. Domain adaptation has been used for a variety of tasks in NLP, including statistical machine translation (Axelrod et al., 2011), sentiment analysis (Blitzer et al., 2007; Glorot et al., 2011) and text classification (Xue et al., 2008).

Recent studies in the field of readability and language proficiency have used a similar approach to the one proposed in this paper. For example, Pilán et al. (2016) tackle the problem of data sparsity when classifying language proficiency levels of learner-written output by incorporating knowledge in the trained model from another domain consisting of input texts written by teaching professionals for learners. Their results indicated that the weighted combination of the two types of data performed best, even when compared to systems based on considerably larger amounts of in-domain data. In this paper we go a step further by applying this approach to readability classification for people with cognitive disabilities.

3 Evaluation of Text Passages by Readers with Autism

We present a collection of 27 individual documents for which the readability was evaluated by 27 different people with a formal diagnosis of autism. The collection is henceforth referred to as the *ASD Comprehension corpus* and is available at: `https://github.com/victoria-ianeva/ASD-Comprehension-Corpus`. Participants were asked to read text passages and answer three multiple choice questions (MCQs) per passage. Evaluation of the difficulty of the texts was then based on their answers to the questions[2].

Participants The evaluation of the texts was performed in three cycles of data collection and involved 27 different participants with autism. Texts 1-9 and 21-27 were evaluated by Group 1, consisting of 20 adult participants (13 male, 7 female) with mean age in years $\mu = 30.75$ and standard deviation $\sigma = 8.23$, while years spent in education, as a factor influencing reading skills, were $\mu = 15.31$, with $\sigma = 2.9$. Texts 10-17 were evaluated by Group 2, consisting of 18 adult participants (11 male and 7 female) with mean age $\mu = 36.83$, $\sigma = 10.8$ and years spent in education $\mu = 16$, $\sigma = 3.33$. Group 3 evaluated texts 18-20 and consisted of 18 adults (12 male and 6 female) with mean age $\mu = 37.22$, $\sigma = 10.3$ and years spent in education $\mu = 16$, $\sigma = 3.33$. All participants had a confirmed diagnosis of autism

[2]While reading the texts and answering the questions, the eye movements of the participants were recorded using an eye tracker; however, the recorded gaze data was not used in this study, hence we do not report details about the gaze data except when describing the data collection procedure. More details can be found in Yaneva (2016).

and were recruited through 4 local charity organisations. None of the 27 participants had other conditions affecting reading (e.g. dyslexia, intellectual disability, aphasia etc.). All participants were native speakers of English.

Materials A total of 27 text passages of varying complexity were collected from the Web. The registers were miscellaneous, covering educational (7 documents), news (10 documents) and general articles (3 documents), as well as easy-to-read texts (7 documents). The average number of words per text was $\mu = 156$ with standard deviation $\sigma = 49.94$. The texts covered a range of readability levels, where the average was $\mu = 65.07$ with $\sigma = 13.71$ according to the Flesch Reading Ease (FRE) score (Flesch, 1949), which is expressed on a scale from 0 to 100 (the higher the score, the easier the text).

A limitation of the study is the small size of the corpus, which was necessary in order to avoid fatigue in the participants and to comply with ethical considerations. By comparison, LocalNews (Feng, 2009), which is the only other corpus for English whose readability has been evaluated by people with cognitive disabilities contains 11 original and 11 simplified texts.

Design of the Multiple-Choice Questions Since people with ASD are generally known to understand many parts of what they read literally (Happé and Frith, 2006; Happe, 1997; Frith and Snowling, 1983; O'Connor and Klein, 2004), it is of interest to examine different types of comprehension of the texts in the ASD corpus. Impairment in specific types of reading comprehension merits the exploration of readability features related to those specific types. Table 1 shows the main types of comprehension we examine in our study following a taxonomy formulated by Day and Park (2005). The table also shows the relationship between the types of comprehension examined and the reading profile of people with autism.

These types of reading comprehension were examined through the inclusion of three multiple-choice questions per text passage, each of which contained three possible answers. The example below is a question examining the ability to make inferences:

Black peppered moths became more numerous in urban areas because:

a) They were mutants

c) They were camouflaged due to the airborne pollution

d) The airborne pollution blackened the white moths with soot

Apparatus and Procedure All participants were verbally instructed about the purpose and procedure of the experiment and given a participant information sheet. Once they were familiar with the implications of the research, they signed a consent form, verbal instruction was reinforced and demographic data about age, education and diagnosis was collected. Eye tracking data was recorded[3], hence the eye tracker was calibrated by each participant before the start of the experiment. Texts were presented on a 19" LCD monitor. In order to maximise the internal validity of the experiment, the texts were presented in random order to each participant. This controlled for factors such as fatigue or participants becoming accustomed to the types of questions asked. The order of questions after each text was also randomised, so that it would not influence the answers given by the participants. The effects of memory were controlled by having the relevant passage constantly displayed on the screen. Participants could therefore refer to it whenever they were not sure about the information it contained. While the effects of background knowledge could not be eliminated entirely, the selection of texts was made in such a way as to ensure that this effect would be minimised as far as possible. The participants read all texts and answered all questions, taking as many breaks as they requested. At the end of the experiment, participants were debriefed.

Development of the Gold Standard for ASD The 27 texts from the ASD corpus were used for evaluation of the document-level classifiers. They were divided into classes of *easy* and *difficult* texts based on the answers to the multiple choice questions (MCQs). Each text was evaluated by three MCQs and each correct answer was given 1 point, while each incorrect answer was awarded 0 points. Thus, if a participant had answered two out of three questions correctly for a given text, then that text had an answering score of two for this participant. After that, all answering scores for the participants were summed for each text. The texts

[3]The recorded eye tracking data is not examined in this study.

Comprehension	Characteristics (Day and Park, 2005)	Relation to ASD
Literal	Understanding of the straightforward meaning of the text: facts, dates, vocabulary, etc	Readers with ASD have predominantly literal understanding of language (MacKay and Shaw, 2004).
Reorganisation	The ability to combine *explicitly* given information from different parts of the text: *"Maria Kim was born in 1945"*; *"Maria Kim died in 1990"*. How old was Maria Kim when she died?".	Since this type of question is based on literal understanding it could provide insights exclusively into the role of context, the use of which is challenging for people with ASD (O'Connor and Klein, 2004).
Inference	The ability to use two or more pieces of information to arrive at a third piece of information that is *implicit*: *"He rushed off, leaving his bike unchained"* => He left his bicycle vulnerable to theft.	Types of inferences challenging for ASD: Inferring given or presupposed knowledge as well as new or implied knowledge derived from mental state words, bridging inferences, figurative language.

Table 1: Types of comprehension examined and their relation to ASD

were then ranked and partitioned at a threshold into two groups. Application of a Shapiro-Wilk test showed that the data was non-normally distributed and the two groups were thus compared using the non-parametric Wilcoxon Signed Rank test. The results indicated that the two groups of texts were significantly different from one another ($z = -6.091$, $p < 0.0001$). Thus 18 texts were classified as *easy* and 9 texts were classified as *difficult.*

4 Corpora

This section describes the corpora used for training and evaluation of the readability classifiers. We train classifiers on three corpora, presented below: i) **the WeeBit corpus** (Vajjala and Meurers, 2012), a comparatively large generic corpus used in readability research; ii) **the FIRST corpus**, a small corpus containing original and manually simplified texts, a subset of which have been evaluated in terms of readability in experiments involving 100 people with autism (Jordanova et al., 2013) and finally, iii) **a combination of the two**. After that we tested our classifiers by applying them to previously unseen user-evaluated data. These data consist of two corpora, the readability of which has been evaluated by people with autism (**The ASD Comprehension corpus**, presented above), and by people with intellectual disability (**LocalNews corpus** (Feng et al., 2009)).

4.1 The WeeBit Corpus

The WeeBit corpus (Vajjala and Meurers, 2012) contains educational documents obtained from the Weekly Reader[4] and BBC-BiteSize[5] websites and comprises two sub-corpora of the same names. The Weekly Reader is an educational web-newspaper containing fiction, news and science articles. The WeeklyReader is intended for children aged 7-8 (Level 2), 8-9 (Level 3), 9-10 (Level 4) and 9-12 (Senior level). BBC-BiteSize is also an educational site containing articles at 4 levels corresponding to educational key stages (KS) for children between ages 5-7 (KS1), 7-11 (KS2), 11-14 (KS3) and 14-16 (GCSE). The combined WeeBit corpus comprises 5 readability levels corresponding to the Weekly Reader's Level 2, Level 3 and Level 4 and BBC-BiteSize KS4 and GCSE levels. The corpus contains 615 documents per level. The average document length measured in number of sentences is 23.4 sentences at the lowest level and 27.8 sentences at the highest level.

The WeeBit corpus was the most appropriate to use for the purpose of our work due to the fact that it contains educational and generally informative articles and due to its large size relative to other readability corpora for English. Examples of other corpora include Encyclopedia Britannica (Barzilay and Elhadad, 2003) (40 documents), Literacyworks (Petersen and Ostendorf, 2007) (around 200 documents) or the WeeklyReader (Allen, 2009) on its own. An alternative was to use Wikipedia and Simple English Wikipedia[6] as they contain a very large number of articles; however, claims that Simple English Wikipedia articles are more accessible than English Wikipedia articles have been disputed (Xu et al., 2015; Štajner et al., 2012; Yaneva, 2015).

As the primary purpose of our work is to build two-level readability classifiers, we normalized the WeeBit corpus to include texts of only two

[4]http://www.weeklyreader.com/

[5]http://www.bbc.co.uk/education

[6]http://simple.wikipedia.org/wiki/Main Page

readability levels: *easy* and *difficult*. Thus, the *difficult* texts in our corpus were the ones with class labels BitGCSE and BitKS3 (age 11-16) and the *easy* documents were the ones with class labels WRLevel2 and WRLevel3 (age 9 -11). Texts from Weekly Reader Level4 were excluded from the dataset, as they were intended for students aged 9-12, which overlaps with Weekly Reader Level3 (9-10), BitKS2 (7-11), and BitKS3 (11-14). Thus, the remaining data consisted of 1,610 documents divided into two equally sized classes of *easy* and *difficult* documents.

4.2 The FIRST corpus

The FIRST corpus consists of 25 documents of the registers of popular science and literature (13 texts) and newspaper articles (12 texts) (Jordanova et al., 2013). These texts were presented in both their original and simplified forms, so that the corpus contains 25 paired original and simplified documents (50 documents in total). The simplification was performed by 5 experts working with autistic people, who were given ASD-specific text simplification guidelines, specified in (Jordanova et al., 2013), which contains full details of the simplification procedure and the characteristics of the corpus. In addition to the 50 texts contained in that corpus, original and simplified versions of 6 additional texts were produced in accordance with the specified guidelines. These 12 texts were then evaluated on a sample of 100 adults with autism as part of the evaluation method in the EC-funded FIRST project.[7] Statistically significant differences in the levels of comprehension for texts from the two classes are reported (Jordanova et al., 2013). These texts were added to the FIRST corpus, which thus contains 31 original and 31 simplified versions of documents, of which 6 documents per class were evaluated by people with autism.

4.3 LocalNews Corpus

Similar to the ASD Comprehension corpus, the LocalNews corpus (Feng et al., 2009) is used as test data for evaluating the classifiers. The LocalNews corpus consists of 11 original and 11 simplified news stories and is, to the best of our knowledge, the only other resource in English, for which text complexity has been evaluated by people with intellectual disability. The articles were first manually simplified by humans, a process in which long and complex sentences were split and important information contained in complex prepositional phrases was integrated in separate sentences. Lexical simplification included the substitution of rare words with more frequent ones and the deletion of sentences and phrases not closely related to the meaning of the text. The texts were then evaluated by 19 adults with mild intellectual disability, who showed significant differences in their comprehension scores for the two classes of documents (Feng et al., 2009).

[7] FIRST project. [online] available at: http://www.first-asd.eu/[Last accessed: 19/05/2017]

5 Model Training and Evaluation

This section presents the experiments comparing the performance of the different classifiers.

5.1 Algorithms

The document-level classifier was built using supervised learning algorithms implemented in the Weka toolkit (Frank and Witten, 1998). We evaluated a number of algorithms in the WEKA toolkit and selected the two which performed best when evaluated using 10-fold cross validation over the WeeBit corpus (Random Forests) and the FIRST corpus (Bayes Net). The Random Forest algorithm (Breiman, 2001) is a decision tree algorithm which uses multiple random trees to vote for an overall classification of the given input. The Bayes Net classifier is the implementation of a Bayesian Network classifier (Heckerman et al., 1995) available in Weka. Bayesian networks are probabilistic graphical models which were shown to be very successful in domain adaptation problems (for example Finkel and Manning (2009)). For both learning algorithms we used the default values for their parameters as provided by Weka. Although there is scope for tuning of these parameters, we did not have access to enough data to explore this direction.

5.2 Baseline

We use the Flesch-Kincaid Grade Level readability formula (Kincaid et al., 1981) as a baseline for document classification due to the fact that it is one of the best-performing predictors of text difficulty, and has been used as a baseline in other readability estimation models (Vajjala Balakrishna, 2015). The baseline values are computed by using the score of the formula as a single feature in the classification model.

		Random Forests			Bayes Net		
Feature	**Description**	**W**	**F**	**WF**	**W**	**F**	**WF**
1. Long words	Proportion of words with 3 or more syllables	−	+	−	−	−	−
2. Average word length	Average number of syllables, all words	−	+	−	−	−	−
3. Possible senses	Sum of all senses for all words in the text	−	+	−	−	−	−
4. Polysemous words	Words with more than one sense in WordNet	−	+	−	−	−	−
5. Polysemous type ratio	Ratio polysemous word types / all word types	−	+	+	+	−	+
6. Type-token ratio	Total number of types/number of tokens	−	−	−	−	−	−
7. Vocabulary variation	Word types/ common words not in the text	−	−	−	−	−	−
8. Numerical expressions	Number of numerical expressions	−	+	−	−	−	−
9. Infrequent words	Not in 5,000 most freq. words in English	−	+	−	−	−	−
10. Total number of words	Total number of words in the text	−	−	−	−	−	−
11. Dolch-Fry Index	*Fry 1000 Instant Word List/Dolch Word List*	−	−	−	−	−	−
12. Number of passive verbs	Number of passive verbs	−	+	−	−	−	−
13. Agentless passive density	Incidence score of passive voice	−	−	−	−	−	−
14. Negations	Number of negations	+	+	+	+	+	+
15. Negation density	Incidence score of negations	+	−	+	+	−	−
16. Long sentences	Proportion of sentences longer than 15 words	+	+	+	+	−	+
17. Words per sentence	Total words / total sentences	−	+	+	+	+	−
18. Average sentence length	Sentence length in words	−	+	+	+	+	+
19. Number of sentences	Total number of sentences	−	+	−	−	+	−
20. Paragraph index	10 x total paragraphs / total words	−	+	−	−	−	−
21. Semicolons	Number of semicolons	−	+	−	−	+	−
22. Unusual punctuation	Number of occurences of &, %,	+	+	+	+	−	−
23. Comma index	10 x total commas / total words	−	+	−	−	+	−
24. Pronoun Score	Occurence of pron. per 1,000 words	−	+	−	−	−	−
25. Definite description score	Occurence of def. descr. per 1,000 words	−	+	−	−	−	−
26. Illative conjunctions	Number of illative conjunctions	−	+	+	+	−	+
27. Comparative conjunctions	Number of comparative conjunctions	−	+	−	−	−	−
28. Adversative conjunctions	Number of adversative conjunctions	−	+	+	+	−	−
29. Word frequency	Average frequency of words	−	+	−	−	−	−
30. Age of Acquisition (aver.)	AOA norms from the MRC database	+	−	+	+	−	+
31. Familiarity (average)	Familiarity norms from the MRC database	−	+	−	−	−	−
32. Concreteness (average)	Concreteness norms from the MRC database	−	+	−	−	−	−
33. Imagability (average)	Imagability norms from the MRC database	−	+	+	+	−	−
34. 1st pronominal reference	Number of 1st pronominal ref.	−	−	−	−	−	−
35. 2nd pronominal ref.	Number of 2nd pronominal reference	+	−	+	+	−	+
36. ARI	ARI readability formula (Smith et al., 1989)	−	+	−	−	+	−
37. Coleman-Liau	Coleman-Liau formula (Coleman, 1971)	−	+	−	−	−	−
38. Fog Index	Fog Index formula (Gunning, 1952)	+	+	+	+	+	+
39. Lix	Lix readability formula (Anderson, 1983)	−	−	+	+	−	−
40. SMOG	SMOG formula (McLaughlin, 1969)	−	+	−	−	−	−
41. FRE	Flesch Reading Ease (Flesch, 1948)	−	+	−	−	−	−
42. FKGL	Flesch-Kincaid GL (Kincaid et al., 1981)	−	−	−	−	+	−
43. FIRST readability index	FIRST readability ind. (Jordanova et al., 2013)	−	+	−	−	−	−

Table 2: A list of features, their description and their selection for the Random Forests and BayesNet classifiers, where 'W' stands for WeeBit, 'F' stands for FIRST and 'WF' stands for WeeBit + FIRST

5.3 Features and feature selection

A total of 43 features were used in the experiments. Table 2 presents the features, their descriptions, and an indication of whether or not each individual feature was selected for use in the final model of the different readability classifiers. The features used in this study included lexico-semantic (numbers 1 - 14), syntactic (numbers 15-22), cohesion (numbers 23 - 27), and cognitively-motivated features (numbers 28 - 34), as well as 8 readability formulae (numbers 35 - 43) (Table 2). The cohesion and cognitively motivated features were inspired by those used in the Coh-Metrix tool (McNamara et al., 2014). The source for cognitively-motivated features were the word lists in the MRC Psycholinguistic database (Coltheart, 1981), in which each word has an assigned score based on human rankings. The number of personal words in a text is hypothesised to improve ease of comprehension (Freyhoff et al., 1998), which is why evaluation of the *number of first and second person pronominal references* were included as features in the classification model.

Initially, the full-feature sets were used to obtain the baseline models, which were subsequently optimised using the attribute selection filter for su-

Table 3: F Score Results for 10-fold cross validation

	Random Forests			**Bayes Net**		
	Baseline	All features	Selected features	Baseline	All features	Selected features
WeeBit	0.78	**0.988**	0.984	0.838	0.968	**0.978**
FIRST	0.651	0.794	0.825	0.778	0.810	0.841
WeeBit+FIRST	0.77	0.957	0.973	0.831	0.953	0.966

Table 4: F Score Results for the ASD Comprehension corpus and the LocalNews corpus

ASD Comprehension	**Random Forests**			**Bayes Net**		
	Baseline	All features	Selected features	Baseline	All features	Selected features
WeeBit	0.673	**0.927**	0.820	0.667	0.746	0.820
FIRST	0.747	0.782	0.782	0.817	0.782	0.784
WeeBit+FIRST	0.746	0.817	0.855	0.667	0.746	**0.892**
LocalNews	**Random Forests**			**Bayes Net**		
	Baseline	All features	Selected features	Baseline	All features	Selected features
WeeBit	0.818	0.861	**0.954**	0.817	0.908	0.954
FIRST	0.676	0.76	0.705	0.705	0.705	0.760
WeeBit+FIRST	0.818	0.861	0.908	0.817	0.908	**1**

pervised learning which is distributed with Weka (Frank and Witten, 1998) and through iterative elimination of redundant features. This was done at the stage of model evaluation through ten-fold cross validation. The last six columns of Table 2 indicate the lists of selected features for each model. It can be argued that the Random Forest model is already performing a certain degree of feature selection and therefore it may be not necessary to carry out this task on the experiments involving Random Forest. However, analysis of the Random Trees generated by the algorithm revealed that they contain a larger number of features than those selected by our feature selection step. In addition, by performing feature selection we wanted to learn which linguistic features are good indicators of text complexity.

5.4 Evaluation

First, all classifiers were evaluated using 10-fold cross-validation, using the WeeBit, FIRST and WeeBit + FIRST corpora as training sets (Table 3). After that each classifier was tested on previously unseen user-evaluated data. The two sets of unseen data are the ASD Comprehension corpus described in Section 3 and the LocalNews corpus described in Section 4.3. Results for the evaluation on unseen data are presented in Table 4.

For Random Forests we notice that the model trained on the WeeBit corpus performs best when classifying texts from the ASD Comprehension corpus ($F = 0.927$) and from the LocalNews corpus ($F = 0.954$). However, when using the model trained on the Bayes Net algorithm, we see that best external validity for both the ASD Comprehension corpus ($F = 0.892$) and the LocalNews corpus ($F = 1$) is achieved by using the combined WeeBit + FIRST training set.

6 Discussion

In terms of the effects of the size and type of training data used, the results indicate that, in isolation, smaller, population-specific corpora (e.g. FIRST) are not sufficient to achieve optimal classification accuracy; however, in certain cases such as the classification of the LocalNews texts, they do have the potential to boost the performance of a classification model when combined with larger generic corpora ($F = 1$) . Nevertheless, this improvement is subject to choosing a classification algorithm that has optimal performance when trained on the smaller corpus. It is important to note that the most accurate classification of the ASD Comprehension corpus was achieved by training the Random Forests classifier on the WeeBit corpus alone ($F = 0.927$). Hence, the infusion of population-specific and generic corpora is only useful in certain cases, as discussed below. This is in line with results in other fields. For example, Blitzer et al. (2007) investigate domain adaptation for sentiment analysis. Given a pair of source and target domains, they show how it is possible to improve the performance of a sentiment classifier on the target domain when it is trained on data from the source domain with the help of a small annotated corpus from the target domain. However, they show that it is necessary to consider the distance between the two domains as not any pair will lead to good results. For future research, we will con-

sider how it is possible to define a distance metric that can prove useful in our context.

Regarding the effect of the type of the unseen data, we notice that, surprisingly, the pairs of original and simplified articles contained in the LocalNews corpus were predicted 100% correctly by the classifier trained on the combination of texts from WeeBit + FIRST. A possible reason for this is that the introduction of the FIRST corpus together with the larger WeeBit one enables the classifier to capture certain simplification operations (e.g. sentence splitting and lexical simplification) that are common in both LocalNews and FIRST. Achieving such a high score could also have been complemented by the fact that the genre of the documents contained in the LocalNews corpus is closer to the textual genre of the ones of both the WeeBit and of the FIRST corpora. However, this result was only achieved when combining FIRST with the larger WeeBit corpus and was not otherwise replicated by a classifier trained only on the FIRST data. This implies that relatively large data sets are still a prerequisite for the accurate classification of pairs of original and simplified texts. In both cases, when using Random Forests and Bayes Net, a better classification accuracy was achieved for LocalNews ($F = 0.954$ and $F = 1$, respectively) than for the ASD Comprehension corpus ($F = 0.927$ and $F = 0.892$, respectively). This suggests that corpora containing pairs of texts in their original and simplified forms are generally easier to classify than corpora containing only of texts in their original form. This finding has implications for general readability and text simplification research where pairs of texts in their original and manually simplified forms are commonly used for evaluation purposes. In other words, evaluating on such corpora may result in overly optimistic classification results which are less likely to be replicated in a "real-world scenario" with naturally written texts.

The experiments presented above have several limitations. First, the small size of the corpora (a key problem in disability-related research which we target in this article) means that the texts used in this study do not account for the great heterogeneity of natural language. In an attempt to compensate for the small number of texts, we have tried to include documents from miscellaneous registers and with varying levels of readability. Second, both the ASD Comprehension corpus and the LocalNews corpus were evaluated by a relatively small number of participants, which is why individual differences in comprehension may have larger effects on the definition of the gold standard compared to generic readability studies. Nevertheless, as mentioned at the beginning of this article, collecting data from readers with cognitive disabilities is a much needed but challenging task, and the corpora used in this study are currently the only ones of their kind. We contribute to future research in this area by making available the ASD Comprehension corpus.

7 Conclusion

This paper discussed the effects of algorithm selection, training corpora and evaluation corpora for readability research for people with cognitive disabilities, with a view to addressing the problem of the scarcity of user-evaluated data in this setting. First, we presented a collection of 27 individual documents, the readability of which was evaluated by readers with Autism Spectrum Disorder. We then showed that the corpora used for algorithm selection have an effect on the classification performance of the models and that combining large generic readability corpora with small population-specific ones has the potential to boost the classification performance. Finally, we discuss the effects of the type of evaluation data (original articles versus pairs of original and simplified articles) on the classification accuracy and we show that original and simplified documents are easier to classify, and that the combination of generic and population-specific corpora is particularly useful for the classification of such text pairs.

Acknowledgements

This research is part of the AUTOR project partially supported by University Innovation Funds awarded to the University of Wolverhampton.

References

Melissa L Allen. 2009. Brief report: decoding representations: how children with autism understand drawings. *Journal of autism and developmental disorders* 39(3):539–43. https://doi.org/10.1007/s10803-008-0650-y.

Jonathan Anderson. 1983. Lix and rix: Variations on a little-known readability index. *Journal of Reading* 26(6):490–496.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. pages 355–362.

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '03, pages 25–32. https://doi.org/10.3115/1119355.1119359.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic, pages 440–447.

Leo Breiman. 2001. Random forests. *Machine Learning* 45(1):5–32.

E. B. Coleman. 1971. *Developing a technology of written instruction: some determiners of the complexity of prose*, Teachers College Press, Columbia University, New York.

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

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A* 33(4):497–505. https://doi.org/10.1080/14640748108400805.

Richard R. Day and Jeong-Suk Park. 2005. Developing Reading Comprehension Questions. *Reading in a Foreign Language* 17(1).

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read-it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies*. Association for Computational Linguistics, pages 73–83.

William H. Dubay. 2004. *The Principles of Readability*. Impact Information. http://www.impact-information.com/.

Lijun Feng. 2009. Automatic readability assessment for people with intellectual disabilities. *SIGACCESS Access. Comput.* (93):84–91. https://doi.org/10.1145/1531930.1531940.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 229–237.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, pages 276–284.

Jenny Rose Finkel and Christopher D. Manning. 2009. Hierarchical bayesian domain adaptation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. pages 602–610.

R. Flesch. 1949. *The art of readable writing*. Harper, New York.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of applied psychology* 32(3):221–233.

Thomas François. 2015. When readability meets computational linguistics: a new paradigm in readability. *Revue française de linguistique appliquée* 20(2):79–97.

Eibe Frank and Ian H. Witten. 1998. Generating accurate rule sets without global optimization. In J. Shavlik, editor, *Fifteenth International Conference on Machine Learning*. Morgan Kaufmann, pages 144–151.

G. Freyhoff, G. Hess, L. Kerr, B. Tronbacke, and K. Van Der Veken. 1998. Make it simple. european guidelines for the production of easy-to-read information for people with learning disability. Technical report, ILSMH European Association.

Uta Frith and Maggie Snowling. 1983. Reading for meaning and reading for sound in autistic and dyslexic children. *Journal of Developmental Psychology* 1:329–342.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*.

Robert Gunning. 1952. *The technique of clear writing*. McGraw-Hill, New York.

F Happe. 1997. Central coherence and theory of mind in autism: Reading homographs in context. *British Journal of Developmental Psychology* 15:1–12.

Francesca Happé and Uta Frith. 2006. The weak coherence account: Detail focused cognitive style in autism spectrum disorder. *Journal of Autism and Developmental Disorders* 36:5–25.

Francesca GE Happé. 1995. The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child development* 66(3):843–855.

T. L. Harris and R. E. Hodges. 1995. *The Literacy Dictionary: The Vocabulary of Reading and Writing*. International Reading Association.

David Heckerman, Dan Geiger, and David M. Chickering. 1995. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning* 20(3):197–243.

Kentaro Inui, Satomi Yamamoto, and Hiroko Inui. 2001. Corpus-based acquisition of sentence readability ranking models for deaf people. In *NLPRS*. pages 159–166.

Vesna Jordanova, Richard Evans, and Arlinda Cerga Pashoja. 2013. First project - benchmark report (result of piloting task). Central and Northwest London NHS Foundation Trust. London, UK.

J. Peter Kincaid, James A. Aagard, John W. O'Hara, and Larry K. Cottrell. 1981. Computer readability editing system. *IEEE transactions on professional communications* .

Gilbert MacKay and Adrienne Shaw. 2004. A comparative study of figurative language in children with autistic spectrum disorders. *Child Language Teaching and Therapy* 20(13).

Harry G. McLaughlin. 1969. SMOG grading - a new readability formula. *Journal of Reading* pages 639–646.

Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Courtenay Frazier Norbury. 2014. Atypical pragmatic development. *Pragmatic Development in First Language Acquisition* 10:343.

Irene M O'Connor and Perry D Klein. 2004. Exploration of strategies for facilitating the reading comprehension of high-functioning students with autism spectrum disorders. *Journal of autism and developmental disorders* 34(2):115–127.

Sarah E Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *SLaTE*. Citeseer, pages 69–72.

Ildikó Pilán, Elena Volodina, and Torsten Zesch. 2016. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *COLING*. pages 2101–2111.

Luz Rello, Ricardo Baeza-yates, Laura Dempere-marco, and Horacio Saggion. 2012. Frequent Words Improve Readability and Shorter Words Improve Understandability for People with Dyslexia (1):22–24.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*. ACM, New York, NY, USA, CIKM '01, pages 574–576. https://doi.org/10.1145/502585.502695.

Dean R. Smith, A. Jackson Stenner, Ivan Horabin, and III Malbert Smith. 1989. The lexile scale in theory and practice: Final report. Technical report, MetaMetrics (ERIC Document Reproduction Service No. ED307577)., Washington, DC:.

S. Vajjala and D Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 163–173.

Sowmya Vajjala Balakrishna. 2015. *Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications*. Ph.D. thesis, Universität Tübingen.

Sanja Štajner, Richard Evans, Constantin Orasan, and Ruslan Mitkov. 2012. What can readability measures really tell us about text complexity? In Luz Rello and Horacio Saggion, editors, *Proceedings of the LREC'12 Workshop: Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*. European Language Resources Association (ELRA), Istanbul, Turkey.

Sanja Štajner, Ruslan Mitkov, and Gloria Corpas Pastor. 2014. *Simple or not simple? A readability question*, Springer-Verlag, Berlin.

Elisabeth M Whyte, Keith E Nelson, and K Suzanne Scherf. 2014. Idiom, syntax, and advanced theory of mind abilities in children with autism spectrum disorders. *Journal of Speech, Language, and Hearing Research* 57(1):120–130.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics* 3:283–297.

Gui-Rong Xue, Wenyuan Dai, Qiang Yang, and Yong Yu. 2008. Topic-bridged plsa for cross-domain text classification. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '08, pages 627–634.

Victoria Yaneva. 2015. Easy-read documents as a gold standard for evaluation of text simplification output. In *Proceedings of the Student Research Workshop*. INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria, pages 30–36. http://www.aclweb.org/anthology/R15-2005.

Victoria Yaneva. 2016. *Assessing text and web accessibility for people with autism spectrum disorder*. Ph.D. thesis.

Victoria Yaneva and Richard Evans. 2015. Six good predictors of autistic text comprehension. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*. INCOMA

Ltd. Shoumen, BULGARIA, Hissar, Bulgaria, pages 697–706. http://www.aclweb.org/anthology/R15-1089.

Victoria Yaneva, Irina Temnikova, and Ruslan Mitkov. 2015. Accessible texts for autism: An eye-tracking study. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*. ACM, New York, NY, USA, ASSETS '15, pages 49–57. https://doi.org/10.1145/2700648.2809852.