Predicting Reading Difficulty for Readers with Autism Spectrum Disorder

Victoria Yaneva*, Richard Evans*, and Irina Temnikova**
*Research Institute in Information and Language Processing, University of Wolverhampton, UK
**Qatar Computing Research Institute, HBKU, Doha, Qatar
v.yaneva@wlv.ac.uk, r.j.evans@wlv.ac.uk, itemnikova@qf.org.qa

Abstract

People with autism experience various reading comprehension difficulties, which is one explanation for the early school dropout, reduced academic achievement and lower levels of employment in this population. To overcome this issue, content developers who want to make their textbooks, websites or social media accessible to people with autism (and thus for every other user) but who are not necessarily experts in autism, can benefit from tools which are easy to use, which can assess the accessibility of their content, and which are sensitive to the difficulties that autistic people might have when processing texts/websites. In this paper we present a preliminary machine learning readability model for English developed specifically for the needs of adults with autism. We evaluate the model on the ASD corpus, which has been developed specifically for this task and is, so far, the only corpus for which readability for people with autism has been evaluated. The results show that out model outperforms the baseline, which is the widely-used Flesch-Kincaid Grade Level formula.

Keywords: readability, accessibility, autism, automatic text classification

1. Introduction

This paper focuses on the development and evaluation of the first readability model derived by machine learning that is developed specifically for the needs of people with high-functioning autism. Autism Spectrum Disorder (ASD) is a congenital lifelong condition of neural origin, which affects the ability of a person to communicate and interact socially (American Psychiatric Association, 2013).

1.1. Autism Spectrum Disorder

Some people with autism who are less able may remain non-verbal or may develop intellectual disability. People at the higher ends of the spectrum are referred to as high-functioning and are able to produce and comprehend language well, with the exception of certain linguistic constructions and a relative inability to use context and obtain a coherent representation of the text meaning (Happé and Frith, 2006; Frith and Snowling, 1983). At the lexico-semantic level, areas of particular difficulty may include long and unfamiliar words, abstract words and polysemous words, with some autistic people showing deficits in the ability to use context in order to disambiguate homographs (Happé, 1997). At the syntactic level, difficulties include the processing of long sentences containing many clauses, complex punctuation, negation and passive voice, among others (O'Connor and Klein, 2004; Martos et al., 2013). Finally, at the discourse level, readers with autism have been shown to have difficulties grasping the gist of the content of a text as a whole, and difficulties understanding irony, sarcasm, metaphor and authors' intentions (Whyte et al., 2014).

Currently, 1 in 100 people are diagnosed with autism in the UK (Brugha et al., 2012), and it is believed that there are two undiagnosed cases for each one diagnosed (Baron-Cohen et al., 2009). Autism prevalence is expected to increase even more due to recent broadening of the diagnostic criteria and increasing understanding of the characteristics of autism, especially within females. Deficits in reading comprehension are indicated to be one of the reasons for reduced academic achievement and increased school dropout within this population (Brugha et al., 2007).

1.2. Autism and Social Inclusion

Enabling content developers of textbooks, websites, webpages and social media to make their content autism-accessible has the potential to enhance the independence and wellbeing of people with autism, as well as to reduce the resources needed for staff members to support autistic service users in finding relevant information about job accessibility, benefits, disability rights, healthcare, etc. The aim of the readability model presented in this paper is to provide autistic individuals, their tutors, and their carers with an easy way to filter information and to find texts to read that are accessible. The readability model will also provide content developers of websites, textbooks, newspapers, and other media, with an inexpensive, quick, and reliable tool to test the accessibility of their material.

While reading comprehension deficits affect school performance and Web searching behavior, they become an even greater barrier when it comes to using social media such as Twitter, Facebook, WhatsApp, Pinterest, etc. In these environments users need to quickly comprehend written text while chatting and are also exposed to a lot of visual content, which has been shown to affect concentration and comprehension in people with autism (Yaneva et al., 2015). At the same time, social media and the Web are particularly important to people with disabilities because these channels empower them to build an identity in which their disability is not at the forefront, a situation quite different from that in face-to-face communication (especially in the cases of motor, visual or hearing impairments). Communication via social media and the Web allows people with a wide range of disabilities to connect with other people without the complexities of real-world social interactions which have been shown to be especially relevant for those with autism (Bosseler and Massaro, 2003; Putnam and Chong, 2008).

Evidence for the demand of people with autism for accessible and safe social media includes the development of

platforms such as the UK-wide *Autism Connect*[1], in which autistic users can connect to each other in a moderated environment. Accessibility features in these social media include both their simplified design and also their provision of easy-to-read explanations of how to use and navigate the platform. One example of such explanation is the following:

Account - an account is a record of your details. Every user has an account that they have to log in to. The account remembers the things you do and the things that other people have said and done in reply to you [2].

Such accessibility features implemented in disability-friendly social media show how crucial accessible writing is for this population of users.

1.3. Aim of This Study

Two ways in which the demand for accessible writing is currently addressed in English are the *Plain English campaign*[3] and the *Easy-to-read campaign* (Tronbacke, 1997), in which writers follow a set of guidelines to make their text easy to comprehend. In cases where this content is targeted particularly to people with cognitive disabilities, the common practice is to evaluate its complexity via consultations with focus groups of target users, which can be time-consuming and expensive and may also require more than one round of rewriting and evaluation. We aim to address this problem by developing an autism-specific readability assessment model, which can evaluate the accessibility of text content before it has been brought to a focus group for evaluation. The model can also be applied in cases where such groups are not available or the text content is too large to be properly evaluated by humans.
Improving accessibility for users with a certain type of disability may also be of benefit to people with other conditions. This is why, in addition to developing a classifier to distinguish between easy and difficult texts for people with autism, we evaluate the generalizability of this model on a dataset of easy and difficult texts evaluated by people with Mild Intellectual Disability (MID). The main contributions of this research are as follows:

- Development and evaluation of a readability model specifically for people with high-functioning autism
- Development of the ASD corpus, which is a set of reading passages, the complexity of which has been assessed by autistic adults through reading comprehension experiments
- Investigation of the model generalizability on the LocalNews corpus (Feng et al., 2009), containing texts whose complexity has been assessed by readers with mild intellectual disability

To the best of our knowledge, this is the first research to propose a machine-learning based readability classifier for people with autism and is the first study to evaluate an autism-specific readability metric on text passages assessed by autistic users. Furthermore, this classifier is especially relevant to the assessment of Web text content, as the sets of texts used in both training and evaluation were obtained from Web sources.
The rest of this paper is structured as follows. Section 2 discusses related work on readability assessment. Section 3 describes the corpora used for the development of the classifier, including the user-evaluated text passages whose readability was measured in experiments with the participation of autistic readers, and Section 4 presents the linguistic features, specifically matched to the reading difficulties of this population. The training and evaluation of the classifier are presented in Section 5, while Section 6 discusses the implications of this research to the field of accessibility research. The main conclusions and avenues for future work are summarized in Section 7.

2. Related Work

2.1. Readability Assessment

Readability has been defined as the ease of comprehension because of the style of writing (Harris and Hodges, 1995). Other definitions such as the ones by (Pikulski, 1995) add that readability is a construct that takes into account the relationship between specific reader populations, specific texts, and the purpose of reading. Investigations into what makes a text readable and the endeavor to find formal expressions by which to measure it, namely the readability formulae, date as far back as the end of the 19th century (Dubay, 2004) and gained a lot of popularity during the '40s and '50s of the 20th century with the growth of the publishing business. Readability formulae are equations which typically exploit surface features of the text such as word length and sentence length, aiming to predict the difficulty of a text. The most popular readability formulae are the Flesch Reading Ease formula (Flesch, 1948), Flesch-Kincaid Grade Level (Kincaid et al., 1975), Army's Readability Index (ARI) (Senter and Smith, 1967), the Fog Index (Gunning, 1952), the Simple Measure of Gobbledygook (SMOG) (McLaughlin, 1969), etc.
Readability formulae have been criticized for not taking into account features related to the background knowledge and cultural bias of the reader, the way ideas are organized and connected within the text, and the amount of memory and cognitive load imposed by the text on the reader (Benjamin, 2012; Siddharthan, 2006; Dubay, 2004). For instance, while word length has been shown to correlate closely with the lexical difficulty of texts as perceived by their readers (Gunning, 1952), it has been pointed out that this measure does not take into account how abstract or concrete the words of the text are or whether they are truly familiar to readers of a certain age and background. To address these drawbacks, cognitive scientists have developed cognitively-motivated features, which were proposed on the basis of human rankings and which aim to account for the familiarity and age of acquisition of common words, as well as their levels of abstractness, concreteness, imaga-

[1] https://www.autism-connect.org.uk/
[2] https://www.autism-connect.org.uk/index.php/site/siteuse
[3] http://www.plainenglish.co.uk/

bility and meaningfulness (Coltheart, 1981). The majority of these and other cognitively-based lexical features have been computed for a total of 98 538 words and are contained in the MRC Psycholinguistic Database (Coltheart, 1981). These and other cognitively-motivated features such as features of cohesion are implemented in the readability assessment tool Coh-Metrix (McNamara et al., 2010).
Finally, advances in the fields of Natural Language Processing and Artificial Intelligence enable both faster computation of existing statistical features and the development of new NLP-enhanced features which can be used in more complex methods of assessment based on machine learning. This makes large-scale readability assessment feasible and allows customization of the assessment models to specific text content and readership. Examples of this are the unigram models, which have been found particularly suitable for assessment of Web content (Si and Callan, 2001), where the presence of links, email addresses and other elements biases the traditional formulae. Readability assessment for people with different types of cognitive disability has also been investigated and is discussed in the next Section 2.2.

2.2. Readability Assessment for People with Cognitive Disabilities

Individuals with mild intellectual disability have been found to have smaller working memory capacity, resulting in difficulty remembering within- and between-sentence relations (Jansche et al., 2010). Specific readability features developed to capture the characteristics of this particular reader population include entity density (counts of entities such as persons, locations and organisations per sentence) and lexical chains (synonymy or hyponymy relations between nouns) (Jansche et al., 2010; Feng, 2009; Huenerfauth et al., 2009). Evidence from eye-tracking experiments and comprehension questions conducted with Spanish readers with dyslexia, suggests that lexical features such as word length or word frequency are more relevant to people with dyslexia, who do not experience difficulties integrating information from the text but instead struggle with decoding particular letter and syllable combinations (Rello et al., 2012a; Rello et al., 2012b).
Due to the lack of corpora whose reading difficulty levels have been evaluated by people with autism, most readability research for this population has so far been focusing on texts simplified by experts using features matched to reflect the reading difficulties of people with autism (Martos et al., 2013; Štajner et al., 2014; Štajner et al., 2012). User-evaluated texts were used for the first time in an earlier 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, based on reading comprehension experiments (Yaneva and Evans, 2015). The results indicated that 6 features had a high discriminatory power:

1. the number of words per sentence,
2. the number of metaphors per text,
3. the average number of words occurring before the main verb in a sentence,
4. the similarity of syntactic structures of adjacent sentences,
5. the Flesch-Kincaid Grade Level, and
6. the Automated Readability Index.

The current experiment builds upon this work by (1) expanding the set of user-evaluated texts and (2) optimizing combinations of features to distinguish between two classes of difficulty by means of a machine learning algorithm. The process of evaluating the reading passages with people with autism, and the rest of the corpora used for building and evaluating the readability classifier are presented in Section 3.

3. Corpora

The main problem when discussing corpora with respect to training readability classifiers for people with cognitive disabilities is that there is a lack of corpora large enough to be used as a training set. In previous research on readability for people with mild cognitive disability, this issue was addressed by training the classifier on a general corpus with 5 readability levels (The Weekly Reader) and then evaluating it on the LocalNews corpus, a small set of 11 difficult and 11 easy user-evaluated texts (Feng et al., 2009). We propose a similar set-up in which our classifier is trained on the WeeBit corpus (Vajjala and Meurers, 2012), which is a comparatively large corpus consisting of material for schoolchildren of different ages (Section 3.1). After that we evaluate the generalizability of the model on a smaller set of 27 text passages whose difficulty was assessed by adult readers with autism (Section 3.2.1) and on the LocalNews corpus (Feng et al., 2009) (Section 3.2.2), which contains 11 original and 11 simplified versions of newspaper articles whose complexity has been evaluated on readers with mild intellectual disability.

3.1. Training and Intrinsic Evaluation

The WeeBit corpus (Vajjala and Meurers, 2012) comprises two sub-corpora, The Weekly Reader[4] and BBC-BiteSize[5], obtained from educational websites of the same names. The Weekly Reader is an educational web-newspaper with articles from the domains of fiction, news and science intended for children of ages 7-8 (Level 2), 8-9 (Level 3), 9-10 (Level 4) and 9-12 (Senior level). BBC-BiteSize contains 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). After removing audio files and non-textual information (including all of KS1, as it consists mostly of images), 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 with average document length of 23.4 sentences at the lowest level and 27.8 sentences at the highest level.
As the primary purpose of our work is to build a readability classifier for people with autism, we normalized the

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

WeeBit corpus to include texts of only two readability levels: Easy and Difficult, to match the format of the corpus evaluated by people with autism. Thus, texts in the WeeBit corpus with class labels BitGCSE and BitKS3 (age 11-16) were mapped to Difficult and those with class labels WRLevel2 and WRLevel3 (age 9 -11) were mapped to Easy. Instances representing texts of class label Weekly Reader Level4 were filtered from the dataset, as the intended readership of this class (people aged 9-12) overlaps with that of Weekly Reader Level3 (9-10), BitKS2 (7-11), and BitKS3 (11-14).

3.2. Extrinsic Evaluation

3.2.1. ASD Corpus: Developing Reading Passages Evaluated by People with Autism

This section presents the design and procedure for the evaluation of the text complexity of reading passages by people with autism. 27 texts from various domains were evaluated by 26 different people with autism (texts 1-16 by 20 people and texts 17-27 by 18 people).

Design: The participants were asked to read text passages and answer three multiple choice questions (MCQs) per passage. Evaluation of the difficulty of the texts is then based on their answers to the questions and their reading time scores.

Text passages: The text set included a total of 27 text passages which vary in difficulty and were obtained from the Web covering miscellaneous domains and registers (Table 1). The size of the text set is small because the length of each text and the number of texts presented to each participant was selected with a view to avoid fatigue and to comply with ethical considerations. Table 1 summarises some of the characteristics of the texts included in this study. The Flesch-Kincaid Grade Level (FKGL) is proportional to text difficulty. Conversely, Flesch Reading Ease (FRE) score, which is expressed on a scale from 0 to 100, is inversely proportional to text difficulty.

Participants: The texts presented in this study were evaluated in two consecutive sessions by two groups of participants. Texts 1-16 were evaluated by Group 1, consisting of 20 adult participants (7 female, 13 male). Texts 17-27 were evaluated by Group 2, consisting of 18 adult participants (11 male and 7 female). All participants had a confirmed diagnosis of autism and were recruited through 4 local charity organisations. None of the 26 participants had other conditions affecting reading (e.g. dyslexia, intellectual disability, aphasia etc.). Mean age (m) for Group 1 in years was m = 30.75, with standard deviation SD = 8.23, while years spent in education, as a factor influencing reading skills, were m = 15.31, with SD = 2.9. For Group 2, mean age in years was m = 36.83 , SD = 10.8 and years spent in education were m = 16, SD = 3.33. All participants were native speakers of English.

Text classification results: The numbers of correct and incorrect answers provided by each participant to the questions for each text were recorded, as was the reading time measured in seconds. First, each reading time was divided by the number of words in the text in order to obtain raw reading time score. After that an answering score was obtained by counting the number of correct answers each participant had given to the 3 questions for each text. Thus, if a participant had answered 2 out of 3 questions correctly for Text 1, then Text 1 has an answering score of 2 for this participant. Finally, to capture the relation between reading time and correctness of the answers, each answer score was divided by the raw reading time for the same participant in order to obtain one single score per text. This was done because answering score is proportional to comprehension level (the more correct answers, the easier the text), while reading time is inversely proportional to comprehension level: the longer a participant reads a text, the more difficult that text is for the participants. Thus, texts were classified based on one general index for each participant for each text.

A Shapiro-Wilk test showed that the general text scores are non-normally distributed. A Friedman test was performed, confirming that there were significant differences between scores obtained for different texts ($\chi^2(16) = 55.258$, $p < 0.000$). After that a Wilcoxon Signed Rank test with Holm-Bonferroni correction was used to determine where the differences in text scores are and on this basis the texts were divided into two groups of "Easy" texts (texts 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 17, 18, 24, 25, 26 and 27) and "Difficult" texts (texts 12, 13, 14, 15, 16, 19, 20, 21, 22 and 23). A Friedman test was applied to each group individually, indicating that there were no statistically significant

	Genre	Words	FKGL	Flesch
T1	Easy-read	77	8.16	60.11
T2	Easy-read	96	6.73	67.33
T3	Easy-read	74	2.71	92.54
T4	Easy-read	178	5.52	75.33
T5	Easy-read	77	5.79	70.67
T6	Easy-read	121	1.75	95.00
T7	Easy-read	58	6.63	68.16
T8	Educational	163	4.93	79.548
T9	Educational	178	4.671	80.22
T10	Educational	206	7.577	65.437
T11	Educational	189	9.276	56.758
T12	Newspaper	226	11.983	40.658
T13	Newspaper	160	8.866	59.82
T14	Newspaper	163	8.765	66.657
T15	Newspaper	185	14.678	45.34
T16	Newspaper	188	9.823	58.298
T17	General	108	4.243	82.305
T18	General	141	4.561	79.108
T19	Newspaper	166	10.344	57.859
T20	Educational	209	6.087	70.124
T21	Educational	151	5.783	60.258
T22	Educational	158	6.102	57.2013
T23	Newspaper	198	13.204	46.481
T24	General	147	11.035	51.965
T25	Encyclopedic	101	8.229	55.011
T26	Encyclopedic	100	2.943	94.15
T27	Encyclopedic	113	6.963	67.304

Table 1: Characteristics of the texts included in the experiment

differences between the answer scores to the texts in each group (Easy texts: $\chi^2(10) = 15.046$, $p < 0.130$; difficult texts: $\chi^2(5) = 9.676$, $p < 0.085$). A Wilcoxon Signed Rank test confirmed that there was a statistically significant difference between the two groups of Easy and Difficult texts (z = -5. 104, $p < 0.000$).

3.2.2. LocalNews corpus and readers with mild intellectual disability

The LocalNews corpus (Feng et al., 2009) 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 cognitive disabilities. 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 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, showing significant differences between their comprehension scores for the two classes of documents (Feng et al., 2009).

4. Features

A total of 43 features were evaluated in the WeeBit, ASD, and LocalNews corpora. These features are grouped in 5 categories, as presented below.

Lexico-semantic: This group includes surface lexical features such as *Syllables in long words* and *Average word length in syllables*, and semantic features such as *Number of polysemous words* and *Polysemous type ratio*. Lexical diversity is measured through *Type-token ratio, Vocabulary variation* (word types divided by common words not in the text) and *Number of numerical expressions*. Statistical measures include *Numbers of infrequent words*, as well as *Total number of words* and *Dolch-Fry Index*, which evaluates the proportion of words in the text that appear in the *Fry 1000 Instant Word List* (Fry, 2004) or the *Dolch Word List* (Dolch, 1948)

Syntactic: Here were included surface features such as *Long sentences* (proportion of sentences in the text that contain more than 15 words), *Words per sentence* (total words in input file / total sentences in input file), *Average Sentence Length, Total number of sentences* and *Paragraph index* (10 * total paragraphs / total words). Also, features quantifying the number of punctuation marks indicating syntactic complexity were evaluated: *Number of Semicolons/suspension points, Number of Unusual punctuation marks* and *Comma index* (10 * total commas in input file / total words in input file). The cognitive load imposed in syntactic processing by the presence of non-canonical syntactic constructions, verb forms, and modifiers was measured through features such as *Number of passive verbs, Agentless passive density, Negations* and *Negation density*.

Features of cohesion: Cohesion is a property of the text which reflects the ease with which different components are integrated into a whole. As discussed in Section 1, this is especially problematic for readers with autism. We evaluated several features indicating referential and discourse cohesion *Number of illative conjunctions, Comparative conjunctions, Adversative conjunctions, Pronouns* and *Definite descriptions*. These features were computed as in (McNamara et al., 2010)

Cognitively-motivated features: This class of features was obtained through human rankings as explained in Section 2. People with autism have been shown to sometimes find it difficult to form mental representations of word referents if the words are too abstract or unfamiliar (Martos et al., 2013). The source for these features for our classifier were the word lists in the MRC Psycholinguistic database (Coltheart, 1981), where each word has an assigned score as described in Section 2. These features included *Absolute Average Word Frequency, Age Of Acquisition, Imagability, Concreteness* and *Familiarity*. These indices apply only to those words which were present in the MRC database lists, as opposed to all words in the texts, which is why they are referred to as "found only" in Table 2. The number of personal words in a text is hypothesised to improve ease of comprehension (Freyhoff and Van Der Veken, 1998), which is why evaluation of the number of first and second person pronominal references were included as features in the classification model.

Readability formulae: This list included popular formulae such as ARI (Smith et al., 1989), Coleman-Liau (Coleman, 1971), Fog Index (Gunning, 1952), Lix (Anderson, 1983), SMOG Reading Ease (McLaughlin, 1969), Flesch Reading Ease (Flesch, 1948), Flesch Kincaid Grade Level (Kincaid et al., 1975) and FIRST Readability Index (Jordanova et al., 2013). The latter is given by the formula:

$95.43 - (0.076 \times CI) + (0.201 \times PI) - (0.067 \times SI) - (0.073 \times SLI) - (35.202 \times TTR) - (1.060 \times VV) + (0.778 \times DFI)$

Where CI is Comma Index, PI is Paragraph Index, SI is Syllable Index, SLI is Sentence Length Index, TTR is Type Token Ratio, VV is Vocabulary Variation, and DFI is Dolch-Fry Index. It was developed specifically for people with autism in the EC-funded FIRST project by professional in mental healthcare.

5. Training and Evaluation Results

The partial decision tree (PART) classifier distributed in Weka (Frank and Witten, 1998) was used to derive the decision lists presented in Tables 2 and 3. This partial decision tree served as the text classifier in our experiments.[6] Of the classifiers distributed with *Weka*, PART had best performance in testing. The decision list consists of 14 rules. Of the 43 features tested, 28 are directly exploited by this automatically learned rule set.

The learned classification model classifies a text as *difficult* if evaluation of the features presented in Section 4 reveals that it meets all of the conditions in one or more of the sets presented in Table 2. Similarly, the model classifies a text as *easy* if evaluation of the features presented in Section 4

[6]PART is an iterative learning procedure which works by building a partial C4.5 decision tree (Quinlan, 1993) in each iteration and making the "best" leaf into a rule for inclusion in the model.

reveals that it meets all of the conditions in one or more of the sets presented in Table 3.

Set	Feature	Value
1	Long sentences	> 2
	Age of acquisition found only	> 6.04
	Illative conjunctions	> 1
	Pronoun2Incidence	> 0
	Average sentence length	> 10.97
2	Long sentences	> 4
	Age of acquisition found only	> 5.8
	Pronouns	> 11
3	Age of acquisition found only	> 6.51
	Possible senses	> 1844
	Lix	> 27.1
4	Age of acquisition found only	> 6.34
	Spanish readability index	> 67.876001
	ARI	> 7.9
5	Paragraph index	> 0.565217
	AgeOfAcquisition	> 5.51
	Syllable long words	≤ 0.705882
6	AgeOfAcquisitionFoundOnly	> 6.4
	AgeOfAcquisitionFoundOnly	> 6.73
7	ImagabilityFoundOnly	≤ 395.18
	ConcretenessFoundOnly	≤ 362.94
8	AverageSentenceLength	> 11.27
	FamiliarityFoundOnly	≤ 582.53
9	Illative conjunctions	≤ 9

Table 2: Conditions characterising *difficult* texts

Set	Feature	Value
1	AgeOfAcquisitionFoundOnly	≤ 6.51
	Polysemous type ratio	> 0.609442
	AverageSentenceLength	≤ 16.23
	Long sentences	≤ 5
	Fog	≤ 9
	NegationDensity	≤ 10.13
2	Pronoun2Incidence	≤ 12.5
	AverageWordFrequencyAbs	> 359091.82
	Passive verbs	≤ 4
	Average sentence length	≤ 17.16
	Infrequent words	≤ 116
	Adversative conjunctions	≤ 0
3	Polysemous type ratio	> 0.632075
	FleschKincaidGradeLevel	≤ 10.37
	Comma index	> 0.167131
	Illative conjunctions	≤ 6
4	Long sentences	≤ 3
	Fog	≤ 11.7

Table 3: Conditions characterising *easy* texts

We evaluated the classifier with respect to its ability to label input texts as either *easy* or *difficult* for people with ASD. The test data consisted of the three corpora presented in Section 3. Table 4 displays the f_1-scores achieved by the classifier when processing these texts. The WeeBit corpus was exploited as training data. The f_1-scores achieved by the model in classifying texts from this corpus were obtained via ten-fold cross validation.

The table includes statistics on the accuracy of three different versions of the classifier derived from different feature sets (Column *Feature Selection*). The first version (*All*) exploits all 43 features presented in Section 4. The second version (*FKGL*) exploits just one feature, *Flesch-Kincaid Grade Level*. The third version exploits only the 28 features that are used to condition the rules in the sets derived by the PART classifier (*PART*).

Feature Selection	f_1-score		
	WeeBit	Local News	ASD Corpus
All	0.989	1	0.89
FKGL	0.894	0.829	0.654
Features exploited by PART rulesets	0.990	0.725	0.748

Table 4: Evaluation results of the text classifier for the three collections

Table 4 reveals that *PART* is more accurate than the other models in its classification of texts from the WeeBit corpus, but less accurate when classifying texts of the other two categories. Given that we seek to optmimise the classification of texts in *ASD Corpus*, the classifier exploiting the full set of 43 features is preferred in this context. *All* is more accurate than both *FKGL* and *PART* over texts of both *Local News* and *ASD Corpus* categories.

6. Discussion

The results presented in Section 5 show that the classifier trained on the WeeBit corpus outperforms the widely used Flesch-Kincaid Grade Level (FKGL) formula by achieving an f_1 score of 0.89 when classifying texts, compared to f_1 score of 0.654 for FKGL. There are two interesting observations which could be made based on the results from this study.

The baseline model containing all 43 features performed better than the model including only those features which were retained by the features selection algorithm in PART (PART feature set). In fact, when evaluating by 10 fold cross-validation of the WeeBit corpus, use of the PART feature set achieves slightly better performance. However, the model using only these features does not generalise well to the other text collections. We are not certain of their role in the classification process, but the features in the baseline model which were not included in the PART feature set appear to help the classifier to generalise better.

Readers will note that when classifying texts from the LocalNews corpus when exploiting all features, the classifier worked with perfect accuracy. It should be noted that the number of texts in this set is too small to be considered truly representative of those sought by readers with mild intellectual disability. Classifying a single text incorrectly would reduce f_1 to 0.94. Another reason is the fact that the differences between Easy and Difficult documents in the LocalNews texts have been artificially introduced by manual simplification, in which sentence length and word length have been deliberately shortened. As a result, all formulae and classifiers have an advantage when distinguishing between the two classes. This raises an important issue about the kinds of data used to measure the external validity of readability models. In the best case scenario, this data should consist of documents "in their own

right" rather than texts which are modified versions of other texts. This observation gives additional credit to the result obtained over the ASD corpus, in which Easy texts were not derived from Difficult ones. It would be interesting to test whether original and simplified versions of documents would make a suitable training set for readability classifiers for people with cognitive disabilities, where the simplification has been done with respect to the particular difficulties of the target population. Further, it would be interesting to investigate whether a classifier trained on this type of user-specific data would outperform other classifiers trained on larger scale but generic data.

It is important to note that the findings of this paper and the classifications of the texts from the ASD-corpus are relevant to the population of adults with high-functioning autism and are not necessarily applicable to adults at the lower ends of the spectrum, children, or people with cognitive disabilities other than autism.

7. Conclusions and Future Works

This paper presented work towards the development of a machine learning-based classifier which distinguishes between two levels of difficulty of texts for adults with high-functioning autism. First, the ASD corpus was created containing 27 texts classified as easy or difficult through a reading comprehension experiment involving autistic adults. Then a classifier was trained on the WeeBit corpus containing graded educational materials for children between ages 7-16. The generalizability of the model was tested on the ASD corpus and the LocalNews corpus (evaluated on people with mild intellectual disability), where the presented classifier outperformed the widely-used Flesch-Kincaid Grade Level formula (Kincaid et al., 1975) for both datasets.

Future work involves developing a more fine-grained model to distinguish between 3 levels of difficulty suitable for adults with high-functioning autism, as well as adults with autism and comorbid mild intellectual disability. Another future challenge is the development of a tool to distinguish between easy and difficult sentences for this population, thus optimising future text simplification decisions.

8. Acknowledgements

The authors are indebted to all participants, who took part in the reading comprehension experiments, as well as to Dr. Georgiana Marsic for her valuable help with the extraction of some of the features.

9. Bibliographical References

American Psychiatric Association, . (2013). Diagnostic and Statistical Manual of Mental Disorders (5th ed.).

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

Baron-Cohen, S., Scott, F. J., Allison, C., Williams, J., Bolton, P., Matthews, F. E., and Brayne, C. (2009). Prevalence of autism-spectrum conditions: Uk school-based population study. *The British Journal of Psychiatry*, 194(6):500–509.

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24:1–26.

Bosseler, A. and Massaro, D. W. (2003). Development and evaluation of computer-animated tutor for vocabulary and language learning in children with autism. *Journal of Autism and Developmental Disorders*, 33(6):553–567.

Brugha, T., McManus, S., Meltzer, H., Smith, J., Scoth, F. J., and Purdon, S. (2007). Autism spectrum disorders in adults living in households throughout England. Report from the Adult Psychiatric Morbidity Survey 2007. Technical report, The National Health Service Information Centre for Health and Social Care, London.

Brugha, T. S., Cooper, S. A., and McManus, S. (2012). Estimating the Prevalence of Autism Spectrum Conditions in Adults: Extending the 2007 Adult Psychiatric Morbidity Survey. Technical report, NHS, The Health and Social Care Information Centre., London.

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

Coltheart, M. (1981). The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Dolch, E. W. (1948). *Problems in Reading*. The Garrard Press, Champain, IL.

Dubay, W. H. (2004). *The Principles of Readability*. Impact Information.

Feng, L., Elhadad, N., and Huenerfauth, M. (2009). Cognitively motivated features for readability assessment. In *EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Athens, Greece, March 30 - April 3, 2009*, pages 229–237.

Feng, L. (2009). Automatic readability assessment for people with intellectual disabilities. *SIGACCESS Access. Comput.*, (93):84–91, January.

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

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

Freyhoff, G., H. G. K. L. T. B. and Van Der Veken, K. (1998). Make it Simple. European Guidelines for the Production of Easy-to-Read Information for People with Learning Disability. Technical report, ILSMH European Association.

Frith, U. and Snowling, M. (1983). Reading for meaning and reading for sound in autistic and dyslexic children. *Journal of Developmental Psychology*, 1:329–342.

Fry, E. (2004). *1000 Instant Words: The Most Common Words for Teaching Reading, Writing and Spelling*. Teacher Created Resources.

Gunning, R. (1952). *The technique of clear writing*. McGraw-Hill, New York.

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

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

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

Huenerfauth, M., Feng, L., and Elhadad, N. (2009). Comparing evaluation techniques for text readability software for adults with intellectual disabilities. In *Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility*, Assets '09, pages 3–10, New York, NY, USA. ACM.

Jansche, M., Feng, L., and Huenerfauth, M. (2010). Reading difficulty in adults with intellectual disabilities: Analysis with a hierarchical latent trait model. In *Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '10, pages 277–278, New York, NY, USA. ACM.

Jordanova, V., Evans, R., and Pashoja, A. C. (2013). First project - benchmark report (result of piloting task). Central and Northwest London NHS Foundation Trust. London, UK.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. Technical report, CNTECHTRA Research Branch Report.

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

McLaughlin, H. G. (1969). SMOG grading - a new readability formula. *Journal of Reading*, pages 639–646, May.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., and Graesser, A. C. (2010). Coh-Metrix: Capturing Linguistic Features of Cohesion, May.

O'Connor, I. M. and Klein, P. D. (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.

Pikulski, J. J. (1995). Re a d a b i l i t y.

Putnam, C. and Chong, L. (2008). Software and technologies designed for people with autism: What do users want? In *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility*, Assets '08, pages 3–10, New York, NY, USA. ACM.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Rello, L., Baeza-yates, R., Dempere-marco, L., and Saggion, H. (2012a). Frequent Words Improve Readability and Shorter Words Improve Understandability for People with Dyslexia. (1):22–24.

Rello, L., Bayarri, C., and Gorriz, A. (2012b). What is wrong with this word? dyseggxia: A game for children with dyslexia. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '12, pages 219–220, New York, NY, USA. ACM.

Senter, R. J. and Smith, E. A. (1967). Automated Readability Index. Technical Report AMRL-TR-6620, Wright-Patterson Air Force Base.

Si, L. and Callan, J. (2001). A statistical model for scientific readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, pages 574–576, New York, NY, USA. ACM.

Siddharthan, A. (2006). Syntactic simplification and text cohesion. *Research on Language and Computation*, 4:1:77–109.

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

Tronbacke, B. (1997). Guidelines for Easy-to-Read Materials. Technical report, IFLA, The Hague.

Vajjala, S. and Meurers, D. (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*, pages 163–173, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Štajner, S., Mitkov, R., and Pastor, G. C., (2014). *Simple or not simple? A readability question*. Springer-Verlag, Berlin.

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

Yaneva, V. and Evans, R. (2015). Six good predictors of autistic text comprehension. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 697–706, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Yaneva, V., Temnikova, I., and Mitkov, R. (2015). Accessible texts for autism: An eye-tracking study. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*, ASSETS '15, pages 49–57, New York, NY, USA. ACM.