Effects of Lexical Properties on Viewing Time per Word in Autistic and Neurotypical Readers

Sanja Štajner[1] **Victoria Yaneva**[2] **Ruslan Mitkov**[2] **Simone Paolo Ponzetto**[1]

[1]Data and Web Science Group, University of Mannheim, Germany
{sanja, simone}@informatik.uni-mannheim.de
[2]Research Group in Computational Linguistics, University of Wolverhampton, UK
{v.yaneva, r.mitkov}@wlv.ac.uk

Abstract

Eye tracking studies from the past few decades have shaped the way we think of word complexity and cognitive load: words that are long, rare and ambiguous are more difficult to read. However, online processing techniques have been scarcely applied to investigating the reading difficulties of people with autism and what vocabulary is challenging for them. We present parallel gaze data obtained from adult readers with autism and a control group of neurotypical readers and show that the former required higher cognitive effort to comprehend the texts as evidenced by three gaze-based measures. We divide all words into four classes based on their viewing times for both groups and investigate the relationship between longer viewing times and word length, word frequency, and four cognitively-based measures (word concreteness, familiarity, age of acquisition and imagability).

1 Introduction

Online methodologies such as eye tracking and event-related potentials have been extensively used to investigate word processing among neurotypical readers (Rayner et al., 2012; Dehaene and Cohen, 2011). The idea that the duration of gaze fixations and revisits (go-back fixations to a previously fixated object) could be used as a proxy for measuring cognitive load dates back to the *strong eye-mind hypothesis* by Just and Carpenter (1980), according to which, "there is no appreciable lag between what is fixated and what is processed" (Just and Carpenter, 1980). That is, when a subject looks at something, he/she also processes it cognitively and the amount of time the subject spends on processing the particular object is equal to the amount of time his/her gaze stays fixated on this object. According to this hypothesis, gaze duration metrics allow measuring the cognitive load imposed on the reader by certain words, clauses and sentences (Just and Carpenter, 1980).

A series of studies investigating the effects of word frequency, verb complexity and lexical ambiguity (Juhasz and Rayner, 2003; Rayner et al., 2012), as well as contextual effects on word perception (Ehrlich and Rayner, 1981) concluded that long, rare and ambiguous words are more likely to be fixated longer and their processing requires more cognitive effort from the reader. These are also words that are likely to be replaced with shorter and more frequent ones during lexical simplification aimed at making text more accessible to wider populations (Bott et al., 2012; Glavaš and Štajner, 2015).

Eye tracking has also been extensively used for the investigation of reading-related disorders owing to its capacity to provide information about the online processing of the text. For example, aphasic readers show "qualitatively different gaze fixation patterns" when answering comprehension questions (Dickey et al., 2007) and readers with dyslexia have been found to exhibit longer fixation durations and less efficient scanning techniques (Kim and Lombardino, 2016).

In spite of the decades-long tradition of using gaze data to investigate word processing among neurotypical readers and readers with reading-related disorders, this methodology has been scarcely used to investigate reading among people with Autism Spectrum Disorder (ASD). People with ASD have been shown to experience comprehension difficulties at lexical, syntactic and pragmatic level (Frith and Snowling, 1983; Happe, 1997; O'Connor and Klein, 2004; Happé and Frith, 2006; Whyte et al., 2014) and thus studies

Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pages 271–281
Copenhagen, Denmark, September 8, 2017.

employing online processing techniques have the potential to cast light on the particular linguistic constructions which people with autism find challenging.

1.1 Autism Spectrum Disorder

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterised by impairment in communication and social interaction (American Psychiatric Association, 2013). A majority of children on the spectrum experience language delay, which results in reading comprehension difficulties later on in their lives, such as resolving ambiguity in meaning (Frith and Snowling, 1983; Happé and Frith, 2006), syntax processing of long sentences (Whyte et al., 2014) and identifying pronoun referents (O'Connor and Klein, 2004).

Unlike people with other developmental conditions such as dyslexia, autistic readers are not considered to have deficits in word decoding, successfully applying both lexical (look-and-say) and phonological (grapheme-to-phoneme conversion) strategies for reading words (Frith and Snowling, 1983; Smith Gabig, 2010). This implies that in the case of readers with autism, decoding difficulties are unlikely to be the reason for longer fixation times. However, while decoding skills are considered intact, there is an evidence of semantic deficit in ASD (Henderson et al., 2011; Löfkvist et al., 2014), and more specifically in word comprehension rather than word production (Charman et al., 2003; Luyster et al., 2008). This suggests that a difficulty with accessing and integrating the semantic representation of words could pose higher cognitive load on readers with autism.

This hypothesis is supported through an online measurement of word processing using gaze data. Sansosti et al. (2013) provide evidence for significant differences between the total fixation durations, number of fixations and number of regressions between autistic and non-autistic adolescents while reading individual sentences, suggesting that the reading task imposed an overall heavier cognitive load on the participants from the ASD group.

Brock et al. (2008) also used gaze data[1] and showed that both the ASD and the control participants were able to use context to successfully disambiguate the ambiguous target words. The studies by Sansosti et al. (2013) and Brock et al. (2008) are, to the best of our knowledge, the only two existing studies investigating reading among people with autism using gaze data; we advance this by i) using a larger dataset from a natural reading task as opposed to individual sentences, ii) identifying which words impose heavier cognitive load on the participants and what their lexical properties are.

1.2 Complex Word Identification

Complex Word Identification (CWI) task received high attention only recently, with findings suggesting that using a CWI module at the beginning of a lexical simplification (LS) pipeline significantly improves performances of LS systems (Paetzold and Specia, 2016c) and with the recently organised SemEval-2016 CWI shared task.[2] The goal of the shared task was building CWI systems which would identify challenging words for non-native English speakers. The dataset consisted of sentences (without context), each with one content word (noun, verb, adjective, or adverb) marked as a target word. The training dataset contained 200 sentences, where each target word was annotated by 20 non-native English speakers as 'easy' or 'complex', depending on whether they understood its meaning or not. The participants were asked to mark the word as 'complex' even if they understood the meaning of the sentence as a whole, as long as they did not understand the word itself. The test set consisted of 9,000 sentences, this time each sentence annotated only by one non-native speaker (300 different annotators in total). The main goal of the task was to predict potentially complex words for a non-native English speaker based on the annotations collected from 20 non-native speakers. The analysis of the crowdsourced annotations revealed that 'complex' words are on average shorter, less ambiguous, and less frequent in Simple English Wikipedia[3]. The results of the shared task (Paetzold and Specia, 2016b) showed that the use of features focused only on isolated words and not their context lead to best performing CWI systems. Among many investigated lexical and syntactic features, some of them taking into account the context of the target word and some not, the word frequency of the target word in Simple English Wikipedia was identified as the best

[1] The study by Brock et al. (2008) did not contain gaze data produced during a reading task. Instead, the participants were asked to look at an image on the screen which was either relevant or irrelevant to the target word they were hearing.

[2] http://alt.qcri.org/semeval2016/task11/

[3] https://simple.wikipedia.org

feature (Wróbel, 2016).

In an earlier organised shared task on English Lexical Substitution at SemEval-2012,[4] which had the aim of providing a framework for evaluation of lexical simplification systems, for each given sentence containing one target 'complex' word and four substitution candidates, participating systems were competing in ranking the four given substitution candidates according to their simplicity, i.e. how easy they are to be understood by fluent but non-native English speakers. The best performing system (Jauhar and Specia, 2012) used a combination of collocational features and four psycholinguistic measures extracted from the MRC (Machine Readable Dictionary) Psycholinguistic Database (Coltheart, 1981):

- Concreteness – the level of abstraction associated with the concept a word describes.
- Imageability – the ability of a given word to arouse mental images.
- Familiarity – the frequency of exposure to a word.
- Age of Acquisition – the age at which a given word is appropriated by a speaker.

1.3 Study Aims and Contributions

We advance previous approaches to CWI by focusing on a new, less-studied population of target readers with autism, and by using a more sophisticated approach based on eye tracking data.

In this study, we use parallel gaze data to study the differences in word processing between participants with autism and a control group of neurotypical (non-autistic) participants in a natural reading task. Our aim is to find out which words could potentially be considered challenging for both groups of readers for the purposes of automatic text simplification (ATS) and to explore which lexical properties underpin the differences in word processing. The contributions of this study are as follows.

We first show that in spite of the fact that both groups achieved similar level of reading comprehension, the reading task imposed significantly heavier cognitive load on the participants with autism as measured by three different gaze measures (Section 3). This finding is consistent with the results of Sansosti et al. (2013) (Section 1.1).

[4]https://www.cs.york.ac.uk/semeval-2012/task1/index.html

Next, we identify which particular words (in their specific contexts) impose heavier cognitive load on each group of participants by clustering them as challenging or not, based on viewing time of each participant individually (Section 4.1), and then classifying them into four classes depending on the number of participants who found them challenging (Section 4.2).

Finally, we investigate the lexical properties which underpin the different processing times for the different word classes in two groups of participants, using both statistical (word frequency and length) and cognitively-based (familiarity, age of acquisition, concreteness, and imagability) features. To account for the context in which the words appear, we treat the same word in different contexts as different entries in our clustering and classification tasks, i.e. we are actually clustering and classifying the Areas of Interest (AOIs) and not the words (Section 4.3).

Identifying such lexical properties has both theoretical and practical implications. On one hand, understanding into what makes a word challenging for a reader with autism could inform future writing guidelines for easy-to-read content and the design of exams and test items for students with autism (Elliott et al., 2010). On the other hand, as shown is Section 1.2, the identification of challenging words based on their lexical properties is on a high demand in the field of Natural Language Processing (NLP) for the purpose of automated text simplification for people with autism and other disorders (Martos et al., 2013; Siddharthan, 2014) as well as for non-native speakers.

2 Data Collection

An experimental group of participants with a diagnosis of autism and a control group of non-autistic participants were asked to read 20 texts while their eye movements were recorded by an eye tracker. In order to explore between-group differences in reading patterns, the groups were matched based on their reading comprehension, as follows. It was important to ensure that i) all participants had understood the presented texts at a similar level and ii) that they read for meaning as opposed to simply skimming through the text, which is why they were asked to answer three multiple-choice (MCQ) questions per passage with three possible answers each. The questions assessed both literal

Texts	Group	Participants	Age in years	Years of schooling
1 - 9	ASD	9 (5 male, 4 female)	m = 33, SD = 9.18	m = 15.66, SD = 2.12
1 - 9	Control	9 (5 male, 4 female)	m = 31.33, SD = 7.48	m = 16.88, SD = 1.83
10 - 17	ASD	14 (8 male, 6 female)	m = 37.9 , SD = 9.6	m = 16, SD = 3.77
10 - 17	Control	13 (9 male, 4 female)	m = 33.84, SD = 9.02	m = 18.54, SD = 3.13
18 - 20	ASD	8 (7 male, 1 female)	m = 36.5, SD = 9.78	m = 15.63, SD = 3.74
18 - 20	Control	10 (6 male, 4 female)	m = 31.3, SD = 6.4	m = 18.1, SD = 2.6

Table 1: Mean age and years spent in formal education for the participants whose gaze data was retained

and inferential reading comprehension and were developed following the taxonomy and guidelines of Day and Park 2005. Gaze data from both groups were collected for 3,636 words.

Materials: A total of 20 text passages with varying complexity were obtained from the Web[5]. The registers were miscellaneous, covering educational (7 documents), news (10 documents) and general informational articles (3 documents). Each text passage was self-contained and coherent (did not refer to information given in the rest of the article and could be comprehended independently of it), did not require specific cultural background to be comprehended and did not contain highly specialised terms, unless they were explained within the text. The average number of words per text was 156 with a standard deviation of 49.94 (min = 74 words and max = 242 words). The texts covered a range of readability levels, with an average Flesch Reading Ease score[6] (Flesch, 1948) of 65.07 and a standard deviation (SD) of 13.71 (min = 40.66, max = 95).

Participants: All participants were native speakers of English, had no diagnosed conditions affecting reading (other than autism in the ASD group) and no diagnosed developmental delay. The participants from the two groups had similar age and similar number of years spent in formal education (Table 1). All participants had normal or corrected vision.

The participants with autism had a confirmed clinical diagnosis obtained in the UK after a referral from a general practitioner and based on the the ADOS diagnostic criteria (Gotham et al., 2007). Out of a total of 27 participants in the ASD group, 11 had a diagnosis of ASD and 15 had a diagnosis of Asperger's syndrome (obtained before the introduction of DSM-5 in 2013). Some participants were diagnosed also with depression (four in ASD group; one in control group) and anxiety (six in ASD group).

The gaze recordings were obtained in three cycles of data collection and the 20 text passages were initially read by a total of 27 different people with a formal diagnosis of autism (texts 1-9 by 20 people, texts 10-17 by 18 people and texts 18-20 by 18 people) and by 31 different neurotypical participants (texts 1-9 by 20 people, texts 10-17 by 18 people and texts 18-20 by 14 people). Participants who performed poorly on comprehension testing, had missing or inaccurate gaze data or were unable to calibrate the eye tracker, were subsequently excluded from the study. The final number of participants whose data was retained and analysed were 21 participants with autism and 19 participants without autism.

Apparatus and Procedure: Texts were presented on a 19 LCD monitor. The device used for recording the gaze of the participants was a Gazepoint GP3 video-based eye tracker[7] (60Hz sampling rate and accuracy of 0.5 - 1 degree of visual angle). The eye tracker was calibrated individually for each participant using a 9-point calibration procedure. The distance between each participant and the eye tracker was controlled by a sensor integrated within the Gazepoint software, and was approximately 65 cm. The software randomised both the order of presentation of the texts and the questions pertaining to texts for each participant, to avoid bias. Participants were instructed about the purpose and the procedure of the experiment, signed a consent form and then read all texts and answered all questions, taking breaks if needed. At the end of the experiment, demographic data was collected and participants were debriefed.

Data Post-Processing: Each word in the texts

[5]The data are available at https://github.com/victoria-ianeva/ASD-Comprehension-Corpus. For more information about the data see Yaneva (2016).

[6]Expressed on a scale from 0 to 100 (the higher the score, the easier the text).

[7]Available at: https://www.gazept.com/

was defined as an Area of Interest (AOI) using the in-built Gazepoint analysis software. The output contains three gaze based measures for a total of 3,636 words for each participant separately: Time Viewed (TV) (the time an AOI was viewed, measured in seconds), Fixations (F) (the number of gaze fixations in a given AOI) and Revisits (R) (the number of go-back fixations in a given AOI, after the eyes have left the AOI and have moved to the right). Cognitive load is usually studied through the temporal aspects of the gaze data. In this paper, we identify challenging words by using the late measure of time viewed per word as opposed to early processing measures such as first fixation duration. This is done in order to account for the overall cognitive load rather than the individual stages of visual word recognition.

3 Between-group Differences in Comprehension and Cognitive Effort

In this section we compare the level of comprehension of the two groups, as well as the duration and number of their fixations and revisits for each word for each participant. A chi-square test for independence revealed that there was no statistically significant association between the group type (ASD vs. Control) and the level of comprehension ($\chi^2(1) = 3.442$; $p = 0.064$). Nevertheless, while both groups achieved similar levels of text comprehension, it took significantly more cognitive effort for the ASD participants to comprehend the text, as shown by all three gaze-based measures (Table 2).[8] This means that identification and simplification of words which pose higher cognitive load on readers with autism could potentially reduce the time and effort required for reading a text, completing an exam, etc.

In order to gain some preliminary insights into the between-group differences we examined the box-plots with outliers and extreme values for TV for each of the 20 texts. We observed that the participants with ASD were more heterogeneous than the control group participants in the words that they viewed extremely long. In contrast, within the control group, the words with extreme TV values were similar for most participants, suggesting that the existing differences between the two groups were not merely based on individual differences between the participants.

Statistic	TV (sec)		Fix.		Rev.	
	ASD	Con.	ASD	Con.	ASD	Con.
Mean	0.20	0.16	1.71	1.46	1.22	0.94
SD	0.29	0.21	1.78	1.41	1.88	1.44
Skewness	5.40	2.47	1.91	1.45	2.55	2.26

Table 2: Eye-tracking data statistics

To better understand the reasons behind certain words been viewed so long and differences between the two groups of participants, we took a systematic approach. We classified all words into four classes using the procedure explained in the next section and then explored the lexical properties of each word class and for each group of participants separately.

4 Between-group Differences in Words Found Challenging

Motivated by the need of automatically recognising potentially challenging words (i.e. CWI task) which should then be replaced by their simpler synonyms in the task of automated text simplification, and the need for ranking substitution candidates according to their simplicity for intended reader (Section 1.2), we wanted to classify all AOIs into different classes according to their potential challenge to the intended reader. Taking into account that different readers might find different words challenging, instead of just classifying words into challenging or not, we wanted to have more fine-grained classes depending on how many readers found them challenging. Therefore, we had a two-step procedure:

1. We divided the words into *challenging* and *not challenging*, according to the TV feature, for each reader separately.

2. We divided the words into four classes, depending on how many readers found them challenging.

4.1 Challenging vs. Not Challenging

The division of words into *challenging* and *not challenging* according to the time viewed could be done in different ways, e.g. by finding a cut-off

[8] Differences in means between the fixations of the two groups of participants for each word were found statistically significant on all three gaze measures using the two-tailed t-test for equality of means in independent samples, where equal variances are not assumed (for TV: $t = 19.842$, df= 61652.575, $p = 0.000$ with 95% CI (0.035, 0.042); for F: $t = 20.781$, df= 64963.384, $p = 0.000$ with 95% CI (0.229, 0.277); and for R: $t = 22.666$, df= 63955.256, $p = 0.000$ with 95% CI (0.263, 0.313)).

Group	Mean	SD	Var.	Min.	Max.
ASD	17.68	4.62	21.37	8.94	27.69
Control	19.81	3.04	9.21	15.14	26.12

Table 3: Percentage of words clustered as challenging (per participant-session combination)

point based on the feature distribution and standard deviation, or by using a parameter-free clustering approach. As there have been no previous studies trying to divide words into those two groups according to the time viewed, and thus no evidence on which approach is better, we opted for the second approach which is parameter-free.

We thus clustered the words from 20 texts into two clusters (*challenging* vs. *not challenging*) for each participant-session combination separately by applying the K-Means algorithm in SPSS, taking only into consideration the TV feature. We applied the iterative KMeans algorithm with two clusters (until convergence, i.e. no change in cluster centers). In a few cases, where there was an extreme outlier (extremely long gazed word) in the given participant-session combination, the clustering resulted in two clusters where one cluster contained only the outlier and the other all other words. In such cases, we applied the K-Means with three clusters, which resulted in having one cluster with *not challenging* words, another with *challenging* words, and the third one with the outlier. We then added the outlier to the cluster of *challenging* words and retained the two resulting clusters.

The average percentage of *challenging* AOIs (out of all words read) was lower, on average, within the ASD group than within the Control group (Table 3).[9] Although this might seem contradictory to the overall higher cognitive load (viewing time) in the ASD group, it is actually a result of the significantly stronger skewness of the TV in the ASD group (Table 2); the participants in the ASD group find fewer AOIs challenging, but they focus on them longer.

4.2 Word Classes

In the second step, for each AOI-id and for each group of participants separately, we assigned one

[9]The between-group differences in percentage of words found challenging were statistically significant using the two-tailed t-test for equality of means in independent samples, where equal variances are not assumed ($t = -2.084$; $df = 45.252$; $p = 0.043$ with 95% CI (-4.184,-0.072)).

Class	# words		% words	
	ASD	Control	ASD	Control
NOT	1,845	1,608	54.51%	47.31%
P-CH	1,158	1,344	34.10%	39.54%
CH	381	444	11.26%	13.06%
E-CH	1	3	3e-4%	9e-4%

Table 4: Distribution of classes

of the following four classes:

- **EXTREMELY CHALLENGING (E-CH)** if that AOI-id was clustered as *challenging* for all participants;
- **CHALLENGING (CH)** if that AOI-id was clustered as *challenging* for at least half of the participants (in the case of the texts read by an odd number of participants, the half was the mean value rounded to the lower integer) but not for all;
- **POTENTIALLY CHALLENGING (P-CH)** if that AOI-id was clustered as *challenging* for at least two participants, but less than a half of the participants;
- **NOT CHALLENGING (NOT)** if none of above (i.e. that AOI-id was clustered as *challenging* for one participant at the most).

The number of AOIs found in each class for each group of participants is presented in Table 4. The distribution of AOIs among classes was similar for both groups of participants, while the numbers supported our hypothesis that the participants in the ASD group are more heterogeneous in the AOIs they find challenging (i.e. the AOIs they viewed long), which results in a lower overlap of challenging AOIs among the participants (i.e. the lower number of POTENTIALLY CHALLENGING (P-CH) and CHALLENGING (CH) AOIs than in the Control group).

Extremely challenging words (E-CH) for the Control group were: *conservative*, *Academicians*, and *iconoclasm*, whereas for the ASD group it was only the word *acquaitance*.

4.3 Importance of Context

In order to account for the influence that the context can have on certain word requiring greater cognitive effort, we were classifying AOIs, allowing thus for the same word (but different AOI) to be classified in different classes.

Word	Context	Class
computer	Experts in Namibia are using a **computer** system to identify and track...	CH
computer	Next, they store the photos on a **computer**.	NOT
computer	Whenever a new print is added, the **computer** compares it to all the other prints...	NOT
comes	Secondhand smoke (SHS) **comes** from burning cigarettes, pipes, or cigars.	NOT
comes	... where an excellent music policy **comes** complete with a decent pint of Guinness.	CH

Table 5: Examples of same words placed in different classes depending on their context.

Class	Age of aquisition (AoA)		Familiarity (Fam)	
	ASD	Control	ASD	Control
NOT CHALLENGING (NOT)	235.1 ± 108.7	230.4 ± 107.5	600.5 ± 71.7	602.8 ± 70.3
POTENTIALLY CHALLENGING (P-CH)	331.0 ± 122.0	317.6 ± 122.4	548.0 ± 83.9	555.5 ± 82.9
CHALLENGING (CH)	427.9 ± 114.3	420.0 ± 115.6	489.5 ± 94.4	495.1 ± 97.6
EXTREMELY CHALLENGING (E-CH)	NotFound	604.7 ± 113.5	NotFound	317.2 ± 162.6

Table 6: Age of aquisition and familiarity of the words in different classes (mean ± standard deviation)

Among the total of 3398 AOIs, 1495 were unique words. Out of those 1495, 1048 appeared only once in the whole corpus (20 texts), 224 appeared twice, 187 appeared between three and ten times, while 36 words appeared more than ten times (stop words only).

For each of the two groups of participants, we closely examined all words that appeared more than once searching for those which (appearing in different contexts) were classified in different levels of difficulty, and especially for those that appear in two not-neighbouring levels (e.g. NOT CHALLENGING and CHALLENGING).

In the case of non-autistic readers, out of 347 words which appeared more than once in the presented texts, 175 were placed always in the same level of difficulty (irrespective of their context), 18 of them (which repeated at least three times) were placed in three different classes (three neighbouring classes – NOT CHALLENGING, POTENTIALLY CHALLENGING, and CHALLENGING), whereas six words (*comes*, *won*, *Foxes*, *provides*, *artists*, *computer*) were placed in two non-neighbouring difficulty levels (NOT CHALLENGING and CHALLENGING).

Two examples of the same words (but different AOIs) classified into two non-neighbouring classes are presented in Table 5 together with contexts.

4.4 Analysis of Word Classes

The mean value with the standard deviation of the cognitively-based features (age of acquisition, familiarity, imagability, and concreteness) in each word class are presented in Tables 6 and 7.

Given that the manually created MRC psycholinguistic database (Coltheart, 1981) covered only 4.76% of words in our texts, we used the bootstrapped larger version of it (Paetzold and Specia, 2016a) which covered 95% of the words.[10]

While the cognitively-based features (age of aquisition, familiarity, imagability and concreteness) were obtained from non-ASD college students, we argue that these properties transfer between subject groups. The reason for this is that our participants were all high-functioning (none of them attended a specialised school) and thus they have all been exposed to a similar vocabulary by going through the national curricula. In addition, both groups understood the texts equally well and we did not observe large between-group differences in the correlation of these metrics with the gaze data.

No significant differences between the values obtained for the same word classes between the two groups of participants were observed. However, it is interesting to note that the extremely

[10]The words not covered by the bootstrapped MRC database (Paetzold and Specia, 2016a) were excluded from the analysis.

Class	Imagability (Img)		Concreteness (Con)	
	ASD	Control	ASD	Control
NOT CHALLENGING (NOT)	354.3 ± 90.3	353.3 ± 90.0	322.4 ± 92.9	322.0 ± 92.8
POTENTIALLY CHALLENGING (P-CH)	390.0 ± 97.4	385.3 ± 96.6	360.9 ± 100.6	355.3 ± 99.9
CHALLENGING (CH)	399.7 ± 89.6	396.6 ± 93.3	376.8 ± 90.9	372.1 ± 95.6
EXTREMELY CHALLENGING (E-CH)	NotFound	302.4 ± 87.1	NotFound	333.3 ± 36.4

Table 7: Imagability and Concreteness of the words in different classes (mean ± standard deviation)

Class	Length		SWiki	
	ASD	Control	ASD	Control
NOT CHALLENGING (NOT)	3.6 ± 1.7	3.5 ± 1.7	0.012 ± 0.018	0.012 ± 0.018
POTENTIALLY CHALLENGING (P-CH)	5.6 ± 2.3	5.3 ± 2.3	0.004 ± 0.012	0.005 ± 0.013
CHALLENGING (CH)	7.8 ± 2.3	7.6 ± 2.4	0.001 ± 0.004	0.001 ± 0.004
EXTREMELY CHALLENGING (E-CH)	11.0 ± NA	11.3 ± 1.2	NotFound	1e-5 ± 2e-5

Table 8: Length and frequency of words in different classes (mean ± standard deviation)

challenging words (E-CH) for the Control group had lower imagability and concreteness than the words classified as less challenging (Table 7). Moreover, the imagability and concreteness values seem to have the opposite correlations with the "challenging" classifications; i.e. the average imagability and concreteness values increase from the NOT to the CH groups. These results imply that the imagability and concreteness may not be well correlated with the cognitive load measured as TV.

The mean value with the standard deviation of the statistically-based measures (length in characters and frequency in Simple Wikipedia) in each word class are presented in Table 8. It is interesting to note that the relative word frequencies in Simple Wikipedia had extremely high standard deviations (Table 8), thus implicating that this feature is not the main characteristic of whether the word is challenging or not.

4.5 Correlation of TV and Word Classes with Lexical Complexity Features

Finally, for each group of participants separately, we tested how the time viewed (taking each participant-AOI combination as a separate data point) and word classes are correlated (using the Spearman's rho coefficient) with both statistical and cognitively-based lexical properties of the words (Table 9).

As can be observed, all investigated lexical properties are better correlated with the word classes than with the raw viewing times (TV). This is probably due to the fact that word classes eliminate the influences of individual differences in reading speed among the participants, which dilute the correlations with the TV.

Feature	TV		Classes	
	ASD	Control	ASD	Control
Len (char.)	+0.297	+0.308	+0.563	+0.556
Con	+0.113	+0.116	+0.241	+0.217
Img	+0.103	+0.107	+0.223	+0.206
AoA	+0.252	+0.261	+0.465	+0.479
Fam	−0.231	−0.235	−0.448	−0.433
SWiki	−0.235	−0.242	−0.457	−0.446

Table 9: Correlation (Spearman's rho) of TV and word classes with lexical complexity features (all statistically significant at a 0.001 level of significance)

5 Discussion

We collected parallel gaze data to study the differences in word processing between participants with autism and a control group of neurotypical participants in a natural reading task.

The presented results indicated that even though both groups understood the texts at a similar level, participants with autism had significantly longer viewing times, more fixations and more revisits per word, indicative of heavier cognitive load. Even when individuals on the spectrum appear highly able and achieve comprehension similar to their peers, they put more cognitive effort into do-

ing so. Another possible explanation of this result could be that the pattern of results observed in the ASD readers reflects a different, perhaps more cautious reading strategy rather than reflecting greater cognitive load associated with lexical processing. In other words, it is possible that given the same instructions, readers with ASD are more careful than control participants to ensure that they have read the text thoroughly and understood the sentences completely. Under this alternative, its not that ASD readers are spending more time and making more fixations because reading is challenging, but instead because they are simply reading more cautiously. Whichever one of these interpretations of the result is valid, this finding provides experimental evidence for the need to allow extra time for exams and for rewriting texts in a way that reduces cognitive load. Both of these accommodations are important steps towards the inclusion of students with ASD.

Although the readers with ASD had significantly longer viewing times, they did not fixate long on as many words as the control participants did. Their overall longer viewing times were heavily skewed towards the words they find challenging. This result reveals differences in the reading patterns between the two groups.

Finally, other than word length which is naturally highly correlated with viewing time, the age of acquisition (AoA) seems to be an important factor related to the viewing times of both groups, followed by frequency and familiarity. This result is consistent with Juhasz and Rayner (2003), who also reported that the effect age of acquisition had on fixation duration was above and beyond the effect of word frequency. Furthermore, the large standard deviation in the word frequency implies that this measure is not suitable for choosing alternative words for lexical substitution in text simplification. Based on our data, an improved strategy for lexical simplification would be basing the word substitutes on the age of acquisition or familiarity ratings. Concreteness and imagability were only weakly related to viewing time. There were no between-group differences observed with regards to the importance of lexical features.

Another important conclusion of this study is that the absolute measures such as concreteness and imagability, which were obtained based on rating of individual words, might not be suitable for complex word identification task, as the gaze data showed that the same word presented in different contexts could be identified as both challenging or not.

One limitation of this study is the fact that it explores only the lexical effects on viewing times and does not explore the effect of contextual features. While we acknowledge the high importance of context for the duration of gaze fixations, the focus on the lexical component in the present study allows for future comparisons between lexical and context-based effects on viewing times. Another limitation is the low speed of the eye tracker used for data collection, which reduces the precision of the recordings and does not allow for comparison of early and late gaze features. However, the data used in this study is the only existing resource of its kind to date and it would be interesting to compare the results obtained from this study with future results based on more sophisticated sets of text and gaze features.

6 Conclusion

This paper presented a study investigating which words are found challenging by readers with high-functioning autism and a control group of non-autistic readers based on gaze data from a natural reading task. We fist showed that even though there were no differences between the level of comprehension of the texts between the two groups, the analysis of the gaze data showed that the readers with autism produced significantly more fixations and revisits, as well as longer viewing times per word. We then clustered the viewing times for each participant-session combination and classified the words into four classes of difficulty based on the gaze data. Finally, we investigated the relationship between those classes and cognitively-based features commonly used in text simplification such as age of acquisition, familiarity, imagability, concreteness, and word frequency and length. Our results showed that relying on such absolute measures for the complex word identification task is not always justified because a given word could be perceived as challenging or not based on the surrounding context.

Acknowledgments

This work has been partially supported by the SFB 884 on the Political Economy of Reforms at the University of Mannheim (project C4), funded by the German Research Foundation (DFG) and the

AUTOR project funded by University Innovation Funds (University of Wolverhampton).

References

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

Stefan Bott, Luz Rello, Biljana Drndarevic, and Horacio Saggion. 2012. Can Spanish be simpler? LexSiS: Lexical simplification for Spanish. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. pages 357–374.

Jon Brock, Courtenay Norbury, Shiri Einav, and Kate Nation. 2008. Do individuals with autism process words in context? evidence from language-mediated eye-movements. *Cognition* 108(3):896–904.

Tony Charman, Auriol Drew, Claire Baird, and Gillian Baird. 2003. Measuring early language development in preschool children with autism spectrum disorder using the macarthur communicative development inventory (infant form). *Journal of child language* 30(1):213.

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

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

Stanislas Dehaene and Laurent Cohen. 2011. The unique role of the visual word form area in reading. *Trends in cognitive sciences* 15(6):254–262.

Michael Walsh Dickey, JungWon Janet Choy, and Cynthia K Thompson. 2007. Real-time comprehension of wh-movement in aphasia: Evidence from eyetracking while listening. *Brain and language* 100(1):1–22.

Susan F Ehrlich and Keith Rayner. 1981. Contextual effects on word perception and eye movements during reading. *Journal of verbal learning and verbal behavior* 20(6):641–655.

Stephen N Elliott, Ryan J Kettler, Peter A Beddow, Alexander Kurz, Elizabeth Compton, Dawn McGrath, Charles Bruen, Kent Hinton, Porter Palmer, Michael C Rodriguez, et al. 2010. Effects of using modified items to test students with persistent academic difficulties. *Exceptional Children* 76(4):475–495.

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

Uta Frith and Maggie Snowling. 1983. Reading for meaning and reading for sound in autistic and dyslexic children. *British Journal of Developmental Psychology* 1(4):329–342.

Goran Glavaš and Sanja Štajner. 2015. Simplifying Lexical Simplification: Do We Need Simplified Corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP), Volume 2: Short Papers*. pages 63–68.

Katherine Gotham, Susan Risi, Andrew Pickles, and Catherine Lord. 2007. The autism diagnostic observation schedule: revised algorithms for improved diagnostic validity. *Journal of autism and developmental disorders* 37(4):613–627.

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.

LM Henderson, PJ Clarke, and MJ Snowling. 2011. Accessing and selecting word meaning in autism spectrum disorder. *Journal of Child Psychology and Psychiatry* 52(9):964–973.

Sujay Jauhar and Lucia Specia. 2012. Uow-shef: Simplex lexical simplicity ranking based on contextual and psycholinguistic features. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval-2012)*. pages 477–481.

Barbara J Juhasz and Keith Rayner. 2003. Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29(6):1312.

Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological review* 87(4):329.

Sunjung Kim and Linda J Lombardino. 2016. Simple sentence reading and specific cognitive functions in college students with dyslexia: An eye-tracking study. *Clinical Archives of Communication Disorders* 1(1):48–61.

Ulrika Löfkvist, Ove Almkvist, Björn Lyxell, and Mari Tallberg. 2014. Lexical and semantic ability in groups of children with cochlear implants, language impairment and autism spectrum disorder. *International journal of pediatric otorhinolaryngology* 78(2):253–263.

Rhiannon J Luyster, Mary Beth Kadlec, Alice Carter, and Helen Tager-Flusberg. 2008. Language assessment and development in toddlers with autism spectrum disorders. *Journal of autism and developmental disorders* 38(8):1426–1438.

Juan Martos, Sandra Freire, Ana González, David Gil, Richard Evans, Vesna Jordanova, Arlinda Cerga, Antoneta Shishkova, and Constantin Orasan. 2013.

FIRST Deliverable - User preferences: Updated. Technical Report D2.2, Deletrea, Madrid, Spain.

I.M. O'Connor and P.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).

Gustavo Paetzold and Lucia Specia. 2016a. Inferring psycholinguistic properties of words. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 435–440.

Gustavo Paetzold and Lucia Specia. 2016b. SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California, USA, pages 560–569.

Gustavo Henrique Paetzold and Lucia Specia. 2016c. Benchmarking lexical simplification systems. In *Proceedings of LREC*. pages 3074–3080.

Keith Rayner, Alexander Pollatsek, Jane Ashby, and Charles Clifton Jr. 2012. *Psychology of reading*. Psychology Press.

Frank J Sansosti, Christopher Was, Katherine A Rawson, and Brittany L Remaklus. 2013. Eye movements during processing of text requiring bridging inferences in adolescents with higher functioning autism spectrum disorders: A preliminary investigation. *Research in Autism Spectrum Disorders* 7(12):1535–1542.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics* 165(2):259–298.

Cheryl Smith Gabig. 2010. Phonological awareness and word recognition in reading by children with autism. *Communication Disorders Quarterly* 31(2):67–85.

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.

Krzysztof Wróbel. 2016. PLUJAGH at SemEval-2016 Task 11: Simple System for Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California, USA, pages 953–957.

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