An analysis of eye-movements during reading for the detection of mild cognitive impairment

Kathleen C. Fraser[1], **Kristina Lundholm Fors**[1], **Dimitrios Kokkinakis**[1], **Arto Nordlund**[2]
[1]The Swedish Language Bank, Department of Swedish
[2]Department of Psychiatry and Neurochemistry
University of Gothenburg, Sweden
{kathleen.fraser, kristina.lundholm, dimitrios.kokkinakis, arto.nordlund}@gu.se

Abstract

We present a machine learning analysis of eye-tracking data for the detection of mild cognitive impairment, a decline in cognitive abilities that is associated with an increased risk of developing dementia. We compare two experimental configurations (reading aloud versus reading silently), as well as two methods of combining information from the two trials (concatenation and merging). Additionally, we annotate the words being read with information about their frequency and syntactic category, and use these annotations to generate new features. Ultimately, we are able to distinguish between participants with and without cognitive impairment with up to 86% accuracy.

1 Introduction

As the global population ages, the prevalence of dementia is increasing (Prince et al., 2013). The term "dementia" refers to an atypical and pathological decline in cognitive abilites, encompassing a range of possible underlying causes. Detecting the onset of dementia as early as possible is important for a number of reasons, including timely access to medication and treatment, increasing support for activities of daily living (such as maintaining proper nutrition and hygiene), reducing the individual's engagement in potentially risky activities (e.g. driving an automobile), and giving individuals, families, and caregivers time to prepare (Solomon and Murphy, 2005; Ashford et al., 2006; Calzà et al., 2015).

In this study, we investigate the possibility of using eye-tracking data and machine learning to detect early, subtle signs of cognitive impairment. Previous work has suggested that changes in eye movements while reading do occur in Alzheimer's disease (Lueck et al., 2000; Fernández et al., 2013; Pereira et al., 2014; Biondi et al., 2017). However, our participants do not have a dementia diagnosis; rather, they have been diagnosed with "mild cognitive impairment", meaning they are starting to show very early signs of cognitive decline, and are at an increased risk of developing dementia. We test the relative merits of collecting eye-tracking data while reading silently and aloud, and explore the idea of augmenting eye-tracking features with linguistic information.

We begin by presenting some background information on cognitive and linguistic changes in dementia, and discuss previous work on eye-tracking and natural language processing approaches to detecting cognitive decline. We then explain our experimental set-up, feature extraction, and machine learning pipeline. We present results for reading silently and reading aloud, and discuss the overall implications and interpretation of our results. Finally we acknowledge the limitations of the current work and suggest areas of future research.

2 Background

There are several different types of dementia, Alzheimer's disease (AD) being the most common one. AD typically debuts with symptoms related to executive cognitive functioning and memory, but also included are specific linguistic impairments, primarily related to semantic processing. Mild cognitive impairment (MCI) can be seen as a stage of pre-clinical dementia, and may manifest years before an actual dementia diagnosis. Persons with MCI show symptoms across several cognitive domains, where global cognitive ability, episodic memory, perceptual speed, and executive functioning are most clearly affected. However, the performance of persons with MCI over-

lap greatly with the performance of healthy controls, which highlights the complexity and heterogeneity of the diagnosis (Bäckman et al., 2005).

Taler and Phillips (2008) reviewed the literature on language impairments in MCI and Alzheimer's disease, and found that the linguistic deficits seen in AD are also present in MCI, albeit to a lesser degree. The main deficits are located on the semantic level (for example, difficulty naming pictures or coming up with words from a particular semantic category), whereas there are no clear evidence of problems regarding syntactic processing. Sentence comprehension is typically impaired in persons with MCI, but there is a great degree of individual variation. Previous research suggests that using tasks that include a possibility to analyse temporal measures (such as reaction time) will improve the ability to distinguish between MCI and healthy controls, and may also be useful as a prognostic factor when investigating which subjects with MCI will convert to AD (Taler and Phillips, 2008).

There has been growing interest in applying machine learning techniques to detect mild cognitive impairment from various linguistic data. Roark et al. (2011) measured the complexity and information content of narrative story re-tellings from 37 participants with MCI and 37 healthy controls, and was able to classify the groups with an AUC of 0.73 using these features alone, or 0.86 by combining this information with clinical test scores. Tóth et al. (2015) leveraged acoustic features (including articulation rate, speech rate, utterance length, pause duration, number of pauses, and hesitation rate) to distinguish between 32 participants with MCI and 19 elderly controls with a best accuracy of 80.4%.

Other research has considered the closely related problem of distinguishing dementia patients from controls through automated analysis of speech and language production (Thomas et al., 2005; Pakhomov et al., 2010; Guinn and Habash, 2012; Meilán et al., 2014; Jarrold et al., 2014; Fraser et al., 2016; Rentoumi et al., 2014; Garrard et al., 2014; Prud'hommeaux and Roark, 2015; Yancheva et al., 2015).

In contrast, computational analyses of language processing and comprehension for the goal of detecting cognitive decline are much rarer, possibly because it is more difficult to quantify automatically. Classical studies of language processing in dementia have considered both listening (for example, Rochon et al. 1994; Kempler et al. 1998; Welland et al. 2002) and reading (for example, Patterson et al. 1994; Storandt et al. 1995); here we focus on reading as the input modality. One well-established method for estimating the processing demands during reading comprehension is through eye-tracking. There is a vast literature on eye-tracking in reading which we will not attempt to fully summarize here, but merely introduce some key vocabulary and basic concepts.

When reading, the eye moves through the text in a series of *fixations* and *saccades*. A fixation occurs when the eye temporarily rests on a word. This time is used to process the incoming information, and to plan the next eye movement. Fixations typically last for around 200-300 ms, and are on average slightly longer in oral reading than in silent reading (Rayner, 1998). In between fixations, the eye makes a rapid movement called a *saccade*. Saccades can move the eye forward through the text (a *forward saccade*) or backward (a *saccadic regression* or simply a *regression*). Saccades tend to be around 6-8 characters in size in English (although this is language-dependent; for example, Liversedge et al. (2016) found longer saccades in Finnish and shorter saccades in Chinese), and around 10-15% of saccades in reading are regressions. Both strong and poor readers make regressions, but stronger readers seem to have the ability to accurately direct their eyes back to a difficult or ambiguous passage, whereas weaker readers perform more general back-tracking (Murray and Kennedy, 1988).

Whether a word is fixated on, and for how long, is influenced by a number of word-level and contextual factors. Content words are fixated on approximately 85% of the time, while function words are fixated on only 35% of the time (Rayner, 1998). There is some evidence that word type effects may be even more fine-grained, as work by Barrett et al. (2016) demonstrates the possibility for part-of-speech tagging based on eye-tracking information. The number and duration of fixations is also affected by word frequency (Raney and Rayner, 1995), word predictability in context (Kliegl et al., 2004), the position of the word in the sentence (Rayner et al., 2000), the emotional valence of the word (Scott et al., 2012), and word length (Rayner, 1998).

While sharing several features, silent reading

and reading aloud are believed to potentially differ in some ways. The main division between the two types of reading is related to the access of phonological and semantic representations in the brain. In silent reading, there has been a great deal of discussion on whether the decoding of orthographic information is directly mapped to semantic meaning, or whether letters are mapped to phonemes, which are then connected to semantic meaning. By using a computational approach based on previous research about reading, Harm and Seidenberg (2004) investigate the two proposed routes and suggest a combined model, where the phonological path and direct path are simultaneously activated and share the workload depending on factors such as word frequency and spelling-sound consistency. The activation of semantic information during reading aloud is also a matter that has been discussed and researched for some time. It was previously thought that during reading aloud, the semantic level of information did not need to be activated, but rather letters could be matched directly to phonemes and then articulated. However, computational models (Coltheart et al., 2001) and for example fMRI data (Graves et al., 2010) have shown that semantic processing is involved in reading aloud, but to varying degrees.

Previous work has identified differences between the eye-movements of individuals with cognitive impairment relative to healthy controls. Lueck et al. (2000) reported that participants with AD had more irregular eye movements when reading, longer fixation times, and more saccadic regressions. Fernández et al. (2013) found that participants with AD had an increased number of fixations and regressions, and also skipped more words than healthy controls. Pereira et al. (2014) presented a review of the literature on eye-tracking in MCI and AD, and suggested that such techniques may be able to predict the conversion from MCI to AD, partly due to the sensitivity of eye-movements to early changes in memory, visual, and executive processes.

Earlier this year, in a paper posted on arXiv, Biondi et al. (2017) reported a classification accuracy of 88.3% in distinguishing between participants with AD and healthy controls through eye-tracking measures. They recorded eye movements from 40 healthy elderly adults and 20 AD patients while they read 120 sentences. The sentences varied in terms of predictability and familiarity (for example, some of the sentences were well-known proverbs). Each sentence was recorded as a separate trial. After removing 10% of the trials as outliers, 90% of the remaining trials were used to train a deep sparse-autoencoder, and 10% were reserved as test data. It is assumed that some of the training data and test data originated from the same participants.

In this paper, we first aim to reproduce aspects of the Biondi et al. (2017) study, although with some notable differences. Our study was conducted in Swedish, rather than Spanish, and in each trial the participant was presented with an entire paragraph, rather than individual sentences, which affects our feature calculations and choice of classifiers. Additionally, we present a comparison of two different trial configurations (reading silently versus reading aloud), and introduce new word-level features to associate linguistic information with the eye-tracking features. Furthermore, perhaps the most critical difference from a clinical standpoint is that our participants are in a milder stage of cognitive decline, and have not received AD diagnoses. Thus we aim explore whether this promising approach can be used to detect the earliest stages of cognitive impairment.

3 Methods

3.1 Participants

The participants were recruited from the Gothenburg MCI study, which is a large longitudinal study on mild cognitive impairment (Wallin et al., 2016). The overall Gothenburg MCI study is approved by the local ethical committee review board (reference number: L09199, 1999; T479-11, 2011); while the currently described study was approved by the local ethical committee decision 206-16, 2016.

To be included in this study, the participants had to fulfill certain inclusion and exclusion criteria: participants had to be native Swedish speakers and had to be able to read and understand information about the project, and be able to give consent. Participants could not have dyslexia or other reading difficulties not relating to their current cognitive impairment. We also excluded patients with deep depression, ongoing substance abuse, poor vision that cannot be corrected with glasses or contact lenses, and participants that were diagnosed with other serious psychiatric, neurological or brain-related diseases, such as Parkinson's dis-

	MCI ($n = 27$)	HC ($n = 30$)
Age (years)	70.3 (5.8)	68.0 (7.5)
Education (years)	14.2 (3.6)	13.3 (3.7)
Sex (M/F)	13/14	9/21
MMSE	28.2 (1.3)	29.6 (0.6)

Table 1: Demographic information for participants with mild cognitive impairment (MCI) and healthy controls (HC). Age, education, and Mini-Mental State Exam (MMSE) scores are given in the format: mean (standard deviation). The MMSE is a general test of cognitive status and has a maximum score of 30.

ease, amyotrophic lateral sclerosis, brain tumour or stroke. Three groups of participants took part in the study: persons with mild cognitive impairment (MCI), persons with subjective cognitive impairment (SCI), and healthy controls (HC). Participants have all been assessed with a battery of tests, from neuropsychological examinations to structural MRI, blood tests, and lumbar punctures. The groups analysed and compared in this paper are the MCI group and the control group. Six control participants and five MCI participants were excluded from the current analysis as a result of calibration problems with the eye-tracker (e.g. due to cataracts or eye inflammation).

Participant information can be seen in Table 1. There is no significant difference between the groups on age or education. The controls do have significantly higher Mini-Mental State Exam (MMSE) scores, on average ($p < 0.0001$). However, we note that the average MMSE score for our MCI participants is 28.2 (out of 30), which is considered to be "normal" (Grut et al., 1993). We contrast this with the AD participants in the study by Biondi et al. (2017), who had an average MMSE score of 24.2. In fact, the healthy control participants in that study had an average MMSE of 27.8, very similar to our MCI group. This indicates the subtle nature of the impairment seen in the MCI category.

3.2 Eye-tracking experiments

The eye-tracking experiments were carried out in a quiet lab environment. We used an EyeLink 1000 Desktop Mount with monocular eye-tracking, and used a headrest for head stabilization. Head stabilization provides an increased eye-tracking performance. The sampling rate was set to 1000 Hz.

The participants read two short texts, and after each text they answered five questions about the texts. The first text was read silently, while the second text was read aloud. Both texts were taken from the International Reading Speed Texts (IReST), which is a collection of texts that is available in 17 different languages. They are 146 words long in Swedish, and were developed to be used as an evaluation tool for impairments in vision or reading ability (Trauzettel-Klosinski et al., 2012). We chose to present complete paragraphs (rather than individual sentences) to simulate a more natural reading task, requiring the integration and recollection of information from the beginning through to the end of the paragraph.

Areas of interest (AOIs) were defined in the text, and each word was labeled as a separate AOI. Eye movements, such as saccades and fixations, are then calculated with respect to the predefined AOIs. Fixations occurring outside the AOIs are not considered in this analysis.

The eye-tracker was calibrated for each participant using a 9-point calibration procedure, and drift-corrected between Trial 1 and Trial 2. However, visual inspection of the data revealed a tendency for downward drift, particularly in the second trial. This was corrected manually, where necessary, to the degree agreed upon by two of the authors (K.C.F. and K.L.F.).

3.3 Features

As our baseline, we consider the 13 features presented in Biondi et al. (2017), and summarized in Table 2. Duration and amplitude features were log-transformed before computing the mean and standard deviation (Wotschack, 2009). The first fixation of each trial is discarded, and analysis starts from the second fixation (Holmqvist et al., 2011). As in Biondi et al. (2017), we partition the fixations into 4 categories: first-pass first fixations, later-pass first fixations[1], multi-fixations, and re-fixations. These definitions are given in Table 2, but for the sake of clarity we also present a simple truth table summarizing the four types of fixations in Table 3.

We then augment these baseline features with information about the words in the text, namely their frequency and word type. We first perform basic syntactic and morphological analysis of the

[1] Biondi et al. (2017) refer to these as "unique" fixations, but this terminology could be ambiguous and thus we have avoided it here.

Gaze duration (mean and s.d.) The mean and standard deviation of the length of time spent fixating on a word, averaged over all words in a trial.
Saccade amplitude (mean and s.d.) The mean and standard deviation of the amplitude of the saccades, averaged over all saccades in a trial.
Total fixations The total number of fixations in a trial.
Total first-pass first fixations The total number of first fixations occurring in the first pass of a trial. That is, a first-pass first fixation occurs when it is the first fixation on the given word, and there have been no fixations on any words occurring later in the text.
Total later-pass first fixations The total number of first fixations occurring outside the first-pass of a trial. That is, a later-pass first fixation occurs when it is the first fixation on the given word, but there have already been fixations on words occurring later in the text.
Total multi-fixations The total number of fixations on a word in the first-pass, excluding the first fixation. That is, a multi-fixation occurs when a word is fixated on multiple times in the run which starts with a first-pass first fixation.
Total re-fixations The total number of fixations on a word outside the first pass, excluding the first fixation.
First-pass first fixation duration (mean and s.d.) The mean and standard deviation of the duration of the first-pass first fixations.
Later-pass first fixation duration (mean and s.d.) The mean and standard deviation of the duration of the later-pass first fixations.

Table 2: Eye-movement features.

		Have any later words been visited?	
		No	Yes
Has this word been visited yet?	No	First-pass first fixation	Later-pass first fixation
	Yes	Multi-fixation	Re-fixation

Table 3: Four types of fixations

two texts using the Sparv annotation tool[2] (Borin et al., 2016). Specifically, each word was lemmatized and labeled with its part-of-speech (POS).

We assign a frequency value for each word lemma according to the number of times it occurs (per one million words) in the "Modern" language section of the Korp Swedish language corpus[3], which contained 10.7 billion word tokens at the time of writing (Borin et al., 2012). These frequency values are POS-disambiguated. We then partition the frequency values into *high* and *low* frequencies, with a threshold of 20 occurrences per million words. This threshold was chosen manually by observing the frequency distribution of the words in the two texts. We also partition the POS labels into two categories: *content* words and *function* words. Content words are defined as nouns, verbs, adjectives, and adverbs; everything else is considered to be a function word.

We then define an augmented feature set, hereafter *Biondi+word*, which takes into account these word-level annotations. Specifically, we create new features corresponding to each of the fixation-based baseline features. (The original feature set also includes saccade amplitude, the computation of which is not attached to any one particular word.) When the original feature involves a mean and standard deviation, we compute the ratio of those values computed on the low:high frequency words and the content:function words. To give an example, for "mean gaze duration", we compute the ratio of the mean gaze duration on low frequency words to mean gaze duration on high frequency words, and the ratio of gaze duration on content words to gaze duration on function words. When the original feature is a raw count, we compute a proportion instead. So for "total fixations", we compute the proportion of total fixations which occur on low-frequency words, and the proportion of total fixations which occur on content words. In this way we define 22 new features to augment the

[2] https://spraakbanken.gu.se/eng/research/infrastructure/sparv

[3] https://spraakbanken.gu.se/eng/korp-info

original Biondi set.

Clearly, we expect these new features to be somewhat correlated with each other, since function words tend also be high-frequency words. However, many content words are also labeled as high-frequency in our methodology, such as *bil* (English: *car*) and *potatis* (English: *potato*).

3.4 Classification framework

We consider three classification algorithms: naïve Bayes (NB), support vector machine (SVM), and logistic regression (LR), implemented in WEKA Version 3.9.1 (Hall et al., 2009). Given the small size of our data set, we forego parameter optimization and use the default parameters; i.e., for LR we use a ridge regression parameter of 10^{-8}, and for SVM we use a first degree polynomial kernel and a complexity parameter of 1.0. For feature selection, we use a wrapper method with a NB classifier. We evaluate the classifier using leave-one-out cross validation, in which at every iteration one data point is held out as a test point, and all remaining points are used for feature selection and classifier training. We report the average classification accuracy across folds. For our dataset, the majority class baseline is 52.6%.

4 Results

4.1 Individual trials

We first consider each trial individually, as we expect there may be differences in eye-movements when reading silently (Trial 1) versus reading aloud (Trial 2). The results for each classifier and each feature set for the first trial are given in Table 4a. Using the augmented feature set hurts classification accuracy in all cases, and the best accuracy of 75.4% is achieved using the naïve Bayes classifier and the Biondi feature set.

When using the data from Trial 2 (Table 4b), the augmented feature set again leads to lower accuracies in all cases, and the best result of 66.7% is achieved by the SVM and naïve Bayes classifiers with the Biondi feature set. In every case, we observe that the classification accuracies are the same or worse on Trial 2 compared to Trial 1. That is, we are able to extract less diagnostically-useful information when the participant is reading aloud than when they are reading silently. This makes sense, since reading aloud is a more constrained task: the reader must keep moving forward at a reasonable pace to avoid disruptions in the spoken narrative. This limits the opportunity for the eyes to move freely around the text. Furthermore, in the reading aloud paradigm, the examiner presented the comprehension questions as soon as the participant had reached the end of the text, in contrast to the silent reading paradigm, in which the participants themselves indicated when they were ready for the questions to be displayed.

	SVM	NB	LR
Biondi	66.7	75.4	73.6
Biondi+word	64.9	71.9	68.4

(a) Trial 1: Reading silently

	SVM	NB	LR
Biondi	66.7	66.7	64.9
Biondi+word	63.1	64.9	63.1

(b) Trial 2: Reading aloud

Table 4: Classifier accuracies for individual trials.

4.2 Combining the trials

We now examine whether we can combine information from the two trials to improve classification accuracy. We consider two different methods for combining the data: (1) concatenating the feature vectors from each trial, and (2) computing the features across both trials, as if they are simply two halves of a single trial. The first method has the advantage of preserving any salient differences between the two experimental paradigms (e.g. if a feature is relevant only when reading silently, that signal will remain in the data). The second method, which we will refer to as *merging*, has the benefit of essentially doubling the amount of data used to compute each feature, possibly leading to more accurate estimates.

The results for each combination are given in Table 5. In most cases, the best accuracy is achieved using the Biondi feature set alone. However, the highest accuracy is 86.0%, which occurs in the merged configuration using the naïve Bayes classifier with the Biondi+word feature set. In every case, a higher accuracy is achieved by merging, rather than concatenating, the data.

4.3 Classification summary

Figure 1 shows the results for each trial and feature set, averaged over the three classifiers. In general, the classifiers trained on Trial 2 did worse than those trained on Trial 1. Concatenating the feature vectors from the two trials resulted in better accu-

	SVM	NB	LR
Biondi	64.9	73.7	68.4
Biondi+word	63.2	73.7	66.7

(a) Concatenated trials

	SVM	NB	LR
Biondi	84.2	82.5	78.9
Biondi+word	84.2	86.0	77.2

(b) Merged trials

Table 5: Classifier accuracies for combined trials.

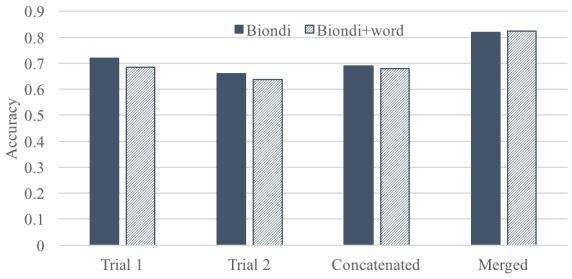


Figure 1: Average accuracies for each trial and feature set, averaged across classifiers.

racies than using Trial 2 data alone, but marginally worse accuracies than using Trial 1 data alone. The best results were achieved by merging the data from the two trials. Using the Biondi feature set alone did better than using the augmented feature set in the first three cases, but the Biondi+word feature set led to slightly higher accuracies in the merged configuration.

However, not all of the observed trends are statistically significant. A 2-way ANOVA revealed a significant effect of trial ($p = 5.0 \times 10^{-7}$) but not feature set on classification accuracy. A Tukey post-hoc test determined that the accuracies in the merged trials are significantly better than in Trial 1 ($p = 6.8 \times 10^{-4}$), Trial 2 ($p = 4.0 \times 10^{-7}$), and the concatenated trials ($p = 1.2 \times 10^{-5}$). However, there is no significant difference between Trial 1 and Trial 2, nor between either of those trials and the concatenated trials.

4.4 Feature analysis

To determine which features help distinguish between the groups, we perform a two-tailed heteroscedastic t-test on all of the features, with Bonferroni correction for repeated comparisons. For this analysis, we consider data from the merged trials, since they led to the best accuracies. Only two features were found to be significantly different between the groups after correction; these are given in Table 6. Consistent with the classification results, none of the frequency or word type features are significant. The total number of first-pass first fixations is significantly higher in the control group but, in contrast, the number of later-pass first fixations is higher in the MCI group. This suggests that the controls have a greater tendency to read through the text from start-to-finish, while the MCI participants tend to skip over words and then return to them later. An example of these different reading patterns can be seen in Figure 2. While this figure only shows data for two participants, it is interesting to note that there is a qualitatively greater difference on the silent trial (Figure 2a and Figure 2c) than in the reading aloud trial (Figure 2b and Figure 2d).

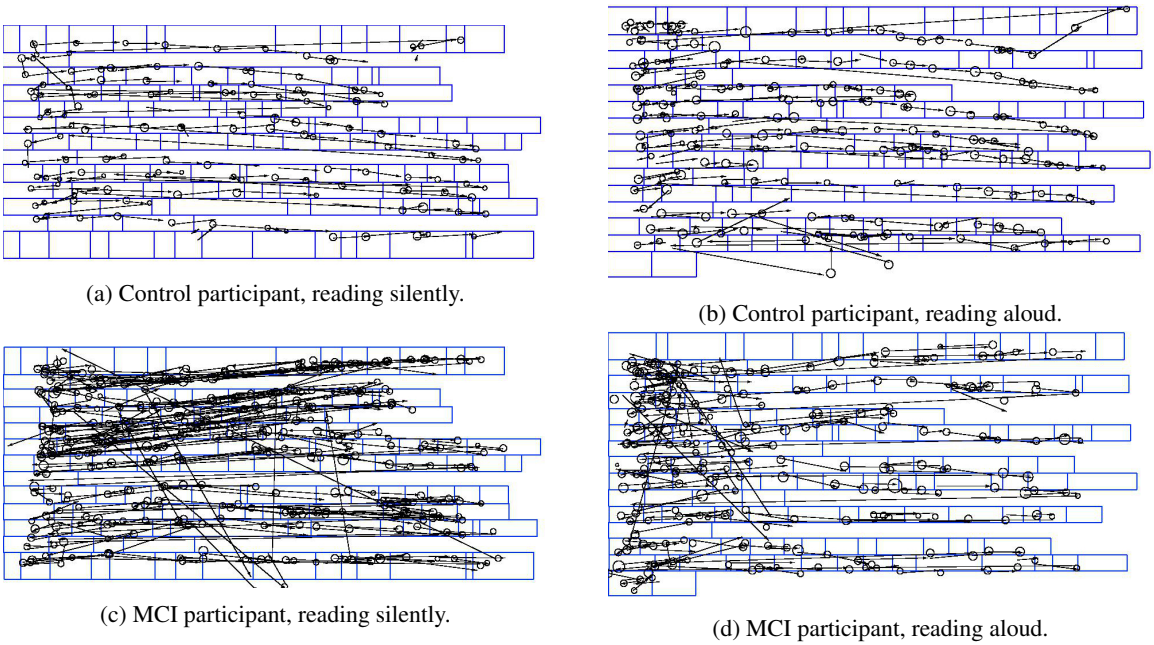
Feature	HC mean	MCI mean	p
First-pass first fixations	98.9	69.1	5.2×10^{-4}
Later-pass first fixations	100.9	133.7	5.8×10^{-6}

Table 6: Features which differ significantly between the groups.

Fernández et al. (2013) found that participants with AD had an increased number of total fixations, first-pass fixations, and second-pass fixations. However, they noted that the second-pass fixations showed an even more striking increase than first-pass fixations. Our results are consistent with this notable increase in second-pass fixations, but not with the reported increase in first pass fixations. One potential reason for this discrepancy could lie in the definition of "first pass fixations", which in Fernández et al. (2013) is given as "the initial reading consisting of all forward fixations on a word", while second-pass fixations are defined as "re-reading"; it is possible that our later-pass first fixations could be classified as first pass fixations under this framework. Nonetheless, both the Fernandez study and our current results suggest a pattern of skipping and back-tracking that is not seen in the control data.

5 Limitations

In this study, as in many studies involving clinical data, our sample is rather small. Furthermore, the two texts were not particularly difficult to read,



(a) Control participant, reading silently.

(b) Control participant, reading aloud.

(c) MCI participant, reading silently.

(d) MCI participant, reading aloud.

Figure 2: Examples of eye movements from a cognitively healthy participant (top) and an MCI participant (bottom), as they read the text from Trial 1 silently (left) and the text from Trial 2 aloud (right). Each blue box in the figures represents an AOI (i.e. a word in the text); the circles indicate fixations and the lines show the movements of the eye. Figure (a) illustrates an example of a relatively straightforward path through the Text 1, while Figure (c) shows one containing more backtracking and re-reading.

nor did they specifically contain words that might be difficult to people with cognitive impairment (for example, low-frequency words with irregular pronunciation, as in Patterson et al. 1994). Additionally, some data had to be either adjusted or in some cases excluded altogether due to calibration quality.

6 Conclusions and future work

In this analysis, we found that we can use eye-tracking information to distinguish between MCI participants and controls with over 80% accuracy, and up to 86% accuracy in the best case. As expected, this is somewhat lower than the accuracy for distinguishing between controls and AD participants reported in Biondi et al. (2017), but demonstrates that eye-tracking may hold promise as a method for detecting the earliest stages of cognitive decline.

We also found that tracking eye movements while the participant reads silently provides more diagnostic information than when reading aloud. Merging data from the two trial conditions led to a significant increase in classification accuracy, compared to using either trial alone. In the merged data set, significant differences between the participant groups were observed for the number of first-pass first fixations (higher in the control group) and later-pass first fixations (higher in the MCI group), suggesting a somewhat disorganized and non-linear path through the text.

Although annotating fixations with the frequency and syntactic category of the word on which the fixation occurs did ultimately lead to the highest classification accuracy, this improvement was not statistically significant, and none of the augmented features showed a significant difference between the HC and MCI groups. It may be that the participants were too early in their decline (and the texts too linguistically simple) for any effect to be seen, or it could be that these variables are not capturing the most relevant linguistic information. In particular, the features were very coarse, making only a binary distinction between high/low frequency words and function/content words. One avenue for future research will be to design more sophisticated ways of incorporating linguistic information into the eye-tracking model, especially features that take into account context, rather than operating at the single word level.

Another untapped source of information is the acoustic signal in the reading aloud trial. Corre-

lating eye movements with acoustic information, such as pauses, fillers, hesitations, and word errors may provide a more complete representation of cognitive processing while reading. Furthermore, other eye-tracking features in addition to those included in the Biondi study may prove to be more sensitive to early cognitive impairment.

In future work we also plan to explore the connection between eye movements and reading comprehension. Each participant in this study also answered comprehension questions related to the passages they read. Analysing the relationship between different eye movement features and the accuracy of the responses may help us better understand the reading strategies used by healthy and cognitively impaired readers.

Finally, future work will include the subjective-cognitive impairment (SCI) group in the analysis. These participants score normally on neuropsychological tests, and so a reliable method for distinguishing them from healthy controls could help provide an early warning system, even before clinical symptoms develop.

Acknowledgments

This work has received support from *Riksbankens Jubileumsfond* - The Swedish Foundation for Humanities and Social Sciences, through the grant agreement no: NHS 14-1761:1.

References

J. Wesson Ashford, Soo Borson, Ruth O'Hara, Paul Dash, Lori Frank, Philippe Robert, William R. Shankle, Mary C. Tierney, Henry Brodaty, Frederick A. Schmitt, Helena C. Kraemer, and Herman Buschke. 2006. Should older adults be screened for dementia? *Alzheimer's & Dementia*, 2(2):76–85.

Lars Bäckman, Sari Jones, Anna-Karin Berger, Erika Jonsson Laukka, and Brent J Small. 2005. Cognitive impairment in preclinical Alzheimer's disease: A meta-analysis. *Neuropsychology*, 19(4):520–531.

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 579–584.

Juan Biondi, Gerardo Fernandez, Silvia Castro, and Osvaldo Agamenonni. 2017. Eye-movement behavior identification for AD diagnosis. *arXiv preprint arXiv:1702.00837*.

Lars Borin, Markus Forsberg, Martin Hammarstedt, Dan Rosen, Roland Schäfer, and Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *The Sixth Swedish Language Technology Conference (SLTC), Umeå University, 17-18 November*.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp - the corpus infrastructure of Språkbanken. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, pages 474–478.

Laura Calzà, Daniela Beltrami, Gloria Gagliardi, Enrico Ghidoni, Norina Marcello, Rema Rossini-Favretti, and Fabio Tamburini. 2015. Should we screen for cognitive decline and dementia? *Maturitas*, 82(1):28–35.

Max Coltheart, Kathleen Rastle, Conrad Perry, Robyn Langdon, Johannes Ziegler, Sally Andrews, Rita Berndt, Derek Besner, Anne Castles, Veronika Coltheart, Martin Davies, Colin Davis, Ken Forster, Carol Fowler, Ram Frost, Jonathan Harrington, Arthur Jacobs, Sachiko Ki-noshita, Ken Paap, Karalyn Patterson, David Plaut, Karen Smith-Lock, Marcus Taft, Anna Woollams, Marco Zorzi We also thank Alice Coltheart, Robert Mannell, Steve Saunders, and Carl Windhorst. 2001. DRC: A Dual Route Cascaded Model of Visual Word Recognition and Reading Aloud. *Psychological Review*, 108:204–256.

Gerardo Fernández, Pablo Mandolesi, Nora P Rotstein, Oscar Colombo, Osvaldo Agamennoni, and Luis E Politi. 2013. Eye movement alterations during reading in patients with early Alzheimer disease. *Investigative ophthalmology & Visual Science*, 54(13):8345–8352.

Kathleen C Fraser, Jed A Meltzer, and Frank Rudzicz. 2016. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Peter Garrard, Vassiliki Rentoumi, Benno Gesierich, Bruce Miller, and Maria Luisa Gorno-Tempini. 2014. Machine learning approaches to diagnosis and laterality effects in semantic dementia discourse. *Cortex*, 55:122–129.

William W Graves, Rutvik Desai, Colin Humphries, Mark S Seidenberg, and Jeffrey R Binder. 2010. Neural Systems for Reading Aloud: A Multiparametric Approach. *Cerebral Cortex August*, 20:1799–1815.

Michaela Grut, L Fratiglioni, M Viitanen, and B Winblad. 1993. Accuracy of the Mini-Mental Status Examination as a screening test for dementia in a Swedish elderly population. *Acta Neurologica Scandinavica*, 87(4):312–317.

Curry I. Guinn and Anthony Habash. 2012. Language analysis of speakers with dementia of the

Alzheimer's type. In *AAAI Fall Symposium: Artificial Intelligence for Gerontechnology*, pages 8–13.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutmann, and Ian H. Witten. 2009. The WEKA data mining software: An update. *ACM Special Interest Group on Knowledge Discovery and Data Mining Explorations*, 2(1).

Michael W Harm and Mark S Seidenberg. 2004. Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, 111(3):662–720.

Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. 2011. *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

William Jarrold, Bart Peintner, David Wilkins, Dimitra Vergryi, Colleen Richey, Maria Luisa Gorno-Tempini, and Jennifer Ogar. 2014. Aided diagnosis of dementia type through computer-based analysis of spontaneous speech. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*, pages 27–36.

Daniel Kempler, Amit Almor, and Maryellen C MacDonald. 1998. Teasing apart the contribution of memory and language impairments in Alzheimer's disease: An online study of sentence comprehension. *American Journal of Speech-Language Pathology*, 7(1):61–67.

Reinhold Kliegl, Ellen Grabner, Martin Rolfs, and Ralf Engbert. 2004. Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, 16(1-2):262–284.

Simon P Liversedge, Denis Drieghe, Xin Li, Guoli Yan, Xuejun Bai, and Jukka Hyönä. 2016. Universality in eye movements and reading: A trilingual investigation. *Cognition*, 147:1–20.

Kristin L Lueck, Mario F Mendez, and Kent M Perryman. 2000. Eye movement abnormalities during reading in patients with Alzheimer disease. *Cognitive and Behavioral Neurology*, 13(2):77–82.

Juan José G Meilán, Francisco Martínez-Sánchez, Juan Carro, Dolores E López, Lymarie Millian-Morell, and José M Arana. 2014. Speech in Alzheimer's Disease: Can temporal and acoustic parameters discriminate dementia? *Dementia and Geriatric Cognitive Disorders*, 37(5-6):327–334.

Wayne S Murray and Alan Kennedy. 1988. Spatial coding in the processing of anaphor by good and poor readers: Evidence from eye movement analyses. *The Quarterly Journal of Experimental Psychology*, 40(4):693–718.

Serguei V.S. Pakhomov, Glen E. Smith, Susan Marino, Angela Birnbaum, Neill Graff-Radford, Richard Caselli, Bradley Boeve, and David D. Knopman. 2010. A computerized technique to asses language use patterns in patients with frontotemporal dementia. *Journal of Neurolinguistics*, 23:127–144.

Karalyn E Patterson, Naida Graham, and John R Hodges. 1994. Reading in dementia of the Alzheimer type: A preserved ability? *Neuropsychology*, 8(3):395.

Marta LG Pereira, Marina von Zuben A Camargo, Ivan Aprahamian, and Orestes V Forlenza. 2014. Eye movement analysis and cognitive processing: detecting indicators of conversion to Alzheimer's disease. *Neuropsychiatric Disease and Treatment*, 10:1273–1285.

Martin Prince, Renata Bryce, Emiliano Albanese, Anders Wimo, Wagner Ribeiro, and Cleusa P. Ferri. 2013. The global prevalence of dementia: A systematic review and metaanalysis. *Alzheimer's & Dementia*, 9(1):63–75.

Emily Prud'hommeaux and Brian Roark. 2015. Graph-based word alignment for clinical language evaluation. *Computational Linguistics*, 41(4):549–578.

Gary E Raney and Keith Rayner. 1995. Word frequency effects and eye movements during two readings of a text. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 49(2):151.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

Keith Rayner, Gretchen Kambe, and Susan A Duffy. 2000. The effect of clause wrap-up on eye movements during reading. *The Quarterly Journal of Experimental Psychology: Section A*, 53(4):1061–1080.

Vassiliki Rentoumi, Ladan Raoufian, Samrah Ahmed, Celeste A. de Jager, and Peter Garrard. 2014. Features and machine learning classification of connected speech samples from patients with autopsy proven Alzheimer's disease with and without additional vascular pathology. *Journal of Alzheimer's Disease*, 42(S3):S3–S17.

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffery Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2081–2090.

Elizabeth Rochon, Gloria S Waters, and David Caplan. 1994. Sentence comprehension in patients with Alzheimer's disease. *Brain and Language*, 46(2):329–349.

Graham G Scott, Patrick J O'Donnell, and Sara C Sereno. 2012. Emotion words affect eye fixations during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(3):783–792.

Paul R. Solomon and Cynthia A. Murphy. 2005. Should we screen for Alzheimer's disease? *Geriatrics*, 60(11):26–31.

Martha Storandt, Katherine Stone, and Emily LaBarge. 1995. Deficits in reading performance in very mild dementia of the Alzheimer type. *Neuropsychology*, 9(2):174.

Vanessa Taler and Natalie A Phillips. 2008. Language performance in Alzheimer's disease and mild cognitive impairment: a comparative review. *Journal of clinical and experimental neuropsychology*, 30(5):501–556.

Calvin Thomas, Vlado Keselj, Nick Cercone, Kenneth Rockwood, and Elissa Asp. 2005. Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. In *Proceedings of the IEEE International Conference on Mechatronics and Automation*, pages 1569–1574.

László Tóth, Gábor Gosztolya, Veronika Vincze, Ildikó Hoffmann, and Gréta Szatlóczki. 2015. Automatic detection of mild cognitive impairment from spontaneous speech using ASR. In *Proceedings of INTERSPEECH*, pages 2694–2698. ISCA.

Susanne Trauzettel-Klosinski, Klaus Dietz, Dürrwächter U, Sokolov AN, Reinhard J, and Klosinski G. 2012. Standardized Assessment of Reading Performance: The New International Reading Speed Texts IReST. *Investigative Opthalmology & Visual Science*, 53(9):5452.

Anders Wallin, Arto Nordlund, Michael Jonsson, Karin Lind, Åke Edman, Mattias Göthlin, Jacob Stålhammar, Marie Eckerström, Silke Kern, Anne Börjesson-Hanson, Mårten Carlsson, Erik Olsson, Henrik Zetterberg, Kaj Blennow, Johan Svensson, Annika Öhrfelt, Maria Bjerke, Sindre Rolstad, and Carl Eckerström. 2016. The Gothenburg MCI study: Design and distribution of Alzheimer's disease and subcortical vascular disease diagnoses from baseline to 6-year follow-up. *Journal of Cerebral Blood Flow and Metabolism : Official journal of the International Society of Cerebral Blood Flow and Metabolism*, 36(1):114–31.

Richard J Welland, Rosemary Lubinski, and D Jeffery Higginbotham. 2002. Discourse comprehension test performance of elders with dementia of the Alzheimer type. *Journal of Speech, Language, and Hearing Research*, 45(6):1175–1187.

Christiane Wotschack. 2009. *Eye Movements in Reading Strategies: How Reading Strategies Modulate Effects of Distributed Processing and Oculomotor Control*. Ph.D. thesis, University of Postdam.

Maria Yancheva, Kathleen Fraser, and Frank Rudzicz. 2015. Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias. In *Proceedings of the 6th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, pages 134–140.