Predicting misreadings from gaze in children with reading difficulties

Joachim Bingel[1] and **Maria Barrett**[2] and **Sigrid Klerke**[3]
[1] Department of Computer Science, University of Copenhagen, Denmark
[2] Centre for Language Technology, University of Copenhagen, Denmark
[3] EyeJustRead, Copenhagen, Denmark
bingel@di.ku.dk, barrett@hum.ku.dk, sk@eyejustread.com

Abstract

We present the first work on predicting reading mistakes in children with reading difficulties based on eye-tracking data from real-world reading teaching. Our approach employs several linguistic and gaze-based features to inform an ensemble of different classifiers, including multi-task learning models that let us transfer knowledge about individual readers to attain better predictions. Notably, the data we use in this work stems from noisy readings *in the wild*, outside of controlled lab conditions. Our experiments show that despite the noise and despite the small fraction of misreadings, gaze data improves the performance more than any other feature group and our models achieve good performance. We further show that gaze patterns for misread words do not fully generalize across readers, but that we can transfer some knowledge between readers using multitask learning at least in some cases. Applications of our models include partial automation of reading assessment as well as personalized text simplification.

1 Introduction

Reading disabilities are impairments affecting individuals' access to written sources, with downstream effects such as low self-confidence in the classroom and limited access to higher education. Dyslexia, for instance, while being highly prevalent with estimates reaching up to 17.5% of the entire population of the U.S. (Interagency Committee on Learning Disabilities, 1987), often goes undiagnosed, such that unattributed weaknesses in reading comprehension further intimidate affected persons. Due to these severe and broad-ranging impacts of reading difficulties, many governments have implemented early screening tests for dyslexia and other reading difficulties and provide special training and assistance for struggling readers throughout the educational system and into adulthood.

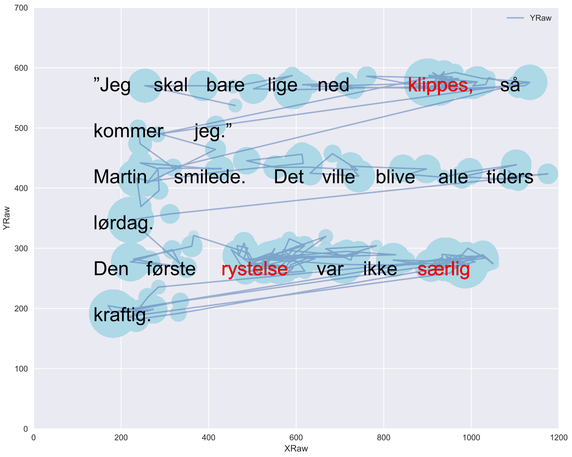


Figure 1: Scanpath and fixations (blue circles) when reading a sentence. This particularly clear example from our dataset shows extended processing time for misread words (marked in red).

In Denmark, for example, such programs provide children with specialist training through focused multi-week reading courses in one-on-one or small group settings. Still, the specialized teachers can only attend to one student at a time when closely monitoring their reading, and the quality of any analysis is strictly limited by the human observer's processing "bandwidth" while attending the live reading.

As a possible mitigation, advances in eye-tracking technology–in particular the increased availability of eye trackers–have made it possible to reliably record children's gaze during reading, both allowing teachers to attend to their students' reading post-hoc as well as providing additional insight into reading strategies based on gaze, including the development of these strategies over time. For the teacher to track and keep records of

Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 24–34
New Orleans, Louisiana, June 5, 2018. ©2018 Association for Computational Linguistics

reading mistakes (henceforth referred to as *misreadings*), however, the students are still required to read out loud, and the teacher has to review the entire reading and annotate for misreadings.

In this work, we investigate to what extent we can predict misreadings from gaze patterns for individual words. While the aim is not to fully automate reading reviews, being able to successfully predict misreadings from gaze data can be part of a semi-automatic system for reading quality assessment and increase teacher efficiency by pointing out potential misreadings for closer review.

Another motivation for this work comes from text simplification, in particular from the observation that individuals' highly specific reading strengths and weaknesses require text simplification models to be customized to specific users in order to unfold their full potential and truly be helpful. Predicting misreadings in concrete reading scenarios and based on individual gaze patterns can be used as a first step in the typical lexical simplification pipeline (Shardlow, 2014).[1] This task, known as complex word identification, has received a considerable amount of attention in the literature, but has exclusively been approached in a user-agnostic fashion.

The data used in this study are gaze recordings of children with reading difficulties, reading Danish texts assigned by their reading teacher as part of their reading intervention. The recordings stem from EyeJustRead, an eye-tracking based software used in special reading intervention in Danish schools.[2] In Section 3, we discuss further aspects of the treatment of gaze data in general and the collection of the data used in this study in particular.

While the difficulty of processing a word is undoubtedly reflected in the fixation time on that word (Rayner et al., 1989), many other factors affect fixation durations, the most prominent being word length and word frequency, but also predictability and relative position in sentence have strong effects–see Figure 1 for a particularly clear example from our dataset. Notably, almost all analyses of eye-tracking reading data use data collected in research laboratories, where these–otherwise confounding–factors can be controlled for. We show that we can perform reasonable misreading detection on real-world eye tracking data, including a limited number of textual features to control for these factors.

Contributions a) We present the first work on the automatic detection of misreadings based on gaze patterns of children with reading difficulties. b) This is, to the best of our knowledge, the first attempt at modeling noisy, real-world eye-tracking data from readers. c) We also present, to the best of our knowledge, the first published results using a multi-task learning setup to transfer knowledge between individual readers for personalized, complex word identification.

2 Related Work

Our work is a special case of complex word identification, a task that has recently received a significant amount of interest, including two shared tasks (Paetzold and Specia, 2016; Yimam et al., 2018). The most successful approaches to these tasks had in common that they employed ensembles of classifiers that learned from a number of semantic and psycholinguistic features. Note however, that these previous approaches to complex word identification aimed at developing generic models that took no account of any specifics of a certain user.

Children's eye movements during reading are not as well-studied as adults', and previous studies typically analyze data collected in experiments designed for research. The overall established observations with regards to reading development are: older children have shorter fixation durations, fewer fixations and fewer regressions. They have a higher skipping probability and also higher saccade amplitude. See Blythe and Joseph (2011) for a review. It is not conclusive whether these variations follow chronological age or their increased reading proficiency. Regardless of the underlying cause, due to the observed systematic differences, the standard procedure is to control as closely as possible for age and reading proficiency level when designing reading experiments.

There are several psycholinguistic studies that show that also in children, the typicality and plausibility of sentences (Joseph et al., 2008) as well as temporary sentence ambiguity (Traxler, 2002) can be traced in eye movements, suggesting that also other types of comprehension difficulties are reflected in the reading patterns.

[1] While today it may hardly sound plausible to equip each laptop with an eye-tracker in order to track people's reading, further technological advances may well make this possible in the future. Recent development in eye-tracking technology has taken it from expensive research equipment to a gaming interface with a price point as low as $100.

[2] http://www.eyejustread.com

Using gaze data to augment models is a recent addition to NLP. Previous approaches that have used gaze data in the context of natural language processing include the work of Barrett et al. (2016), who aim to improve part-of-speech induction with gaze features, Klerke et al. (2016), where gaze data is used as an auxiliary task in sentence compression, and Klerke et al. (2015b), where gaze data is used to evaluate the output of machine translation. The most related work is Klerke et al. (2015a) and Gonzalez-Garduño and Søgaard (2017). Klerke et al. (2015a) compared gaze from reading original, manually compressed, and automatically compressed sentences. They found that the proportion of regressions to previously read text is sensitive to the differences in human- and computer-induced complexity. Gonzalez-Garduño and Søgaard (2017) show that text readability prediction improves significantly from hard parameter sharing when models try to predict word-based gaze features in a multi-task-learning setup. All of these works, however, use gaze data that was collected under laboratory conditions from skilled, adult readers.

3 Gaze Data

In eye-tracking studies, gaze data is normally sampled under experimental circumstances, where e.g. instructions, location, environment, lighting, participant sampling, textual features, order, duration etc. are controlled for. Our real-world data, on the contrary, lacks all of these controls. While in controlled, cognitive psychology experiments, fixation durations have proven to systematically correlate with cognitive load (see Rayner (1998) for a review), eye movements from-real world applications have been largely understudied, and specific findings from the literature on controlled data may not apply here or may be swamped by extraneous factors. Further, the often-used statistical tests of significant differences between gaze patterns lose some of their legitimacy when data is retrieved under noisy conditions.

3.1 Data collection and preprocessing

The data we use in this work is collected in Danish schools using commercial software specifically developed to record and track children's reading development. The system records the eye movements and voice while the children are reading aloud. The teacher can afterwards replay the reading along with the recorded eye movements. The software performs some low-level eye-movement analyses to help the teacher understand how the child processes the text. The teacher can mark which words are erroneously read by the child and later access this and other basic statistics about the reading – see Klerke et al. (2018) for a workflow description. The genre is children's fiction books and the children read contextualized, running text.

As the data is fairly noisy compared to data from laboratory-based eye tracking experiments, we perform thorough cleaning before running any experiments. This cleaning procedure is described below. Table 1 contains a summary of the dataset sizes after each cleaning step. Before any cleaning is performed, the dataset contains 369 reading sessions from 95 unique readers. In total it has 3,161 read pages.

Help word activated on page We start by removing all pages where the reader activated the help word function, which dynamically isolates and enlarges a single word on the screen. This dynamic display generates a series of eye movements that do not resemble typical reading activity. This step removes 94 pages.

Fixation detection We pre-process the raw gaze data by first detecting fixations using a custom implementation of the algorithm of Nyström and Holmqvist (2010). We remove fixations shorter than 40ms and longer than 1.5s.[3] For the calculation of gaze features (see below), we further discard all data points that are not detected as a fixation on text (but instead on images or blank parts of the page). We remove 19 pages where we do not have any fixations on text (e.g. due to the reader just browsing through a book or because of technical issues).

Bad calibration Prior to reading, the student is prompted to calibrate the eye tracker. In the data used in this study, most reading sessions (91%) attain the best calibration score on a five-point scale, while 6% miss a calibration score. The remaining 3% do not have the best calibration score. We remove everything but the 91% with the best calibration score.

Only parts of the readings have been reviewed

[3] Removing short fixations also removes the majority of blinks which presents as a sudden downward-upward pattern of saccades separated by a pause in the signal or a short, falsely detected fixation.

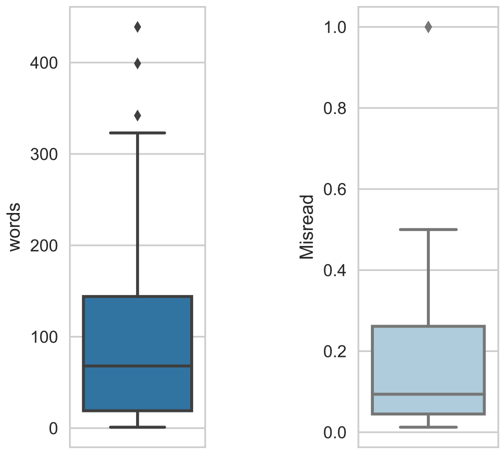
Cleaning step	Reading sessions	Unique readers	Read pages	Read words	Misreadings
No cleaning	369	95	3161	73,965	644
Help word activated	366	95	3067	71,911	619
Fixation detection	366	95	3048	64,191	613
Bad calibration	335	87	2865	56,166	565
Marked by teacher	83	44	405	8,681	565

Table 1: Dataset size after each cleaning step

and marked for misreadings by a teacher. However, whether a teacher reviewed a reading or not is not explicitly encoded in the data. Thus, if there are no marked misreadings in some session, we do not know whether this is because this reading was not reviewed or because there actually were no errors. We therefore remove all readings without any marked misreadings, as well as any data before the first marked misreading and after the last marked misreading within marked sessions, assuming that everything between these two points has been marked. Twelve cleaned reading sessions only consist of one misread word – everything before and after was removed. See Figure 2a for an overview of the distribution of number of words per reading after this cleaning step. This leaves us with the subset of the readings that posed most problems for the subjects. Figure 2b shows the distribution of misread words in the cleaned dataset. It is worth noting that since this is not controlled, experimental data, "misread" is not necessarily interpreted equally by all teachers, or even consistently across markings from the same teacher, due to the lack of an annotation protocol. We assume that "misread" means that the pronounced word deviates substantially from the written word. Ultimately, we retain 83 reading sessions from 44 readers with at least one misread word.

3.1.1 Apparatus

The eye tracker used is a Tobii Eye Tracker 4C with a sample rate of 90 Hz. It is an affordable, consumer eye tracker targeted at gaming. The laptop computers to which the trackers are attached, and which run the software, are provided by the different institutions and vary. Screen resolution is locked by the eye tracker software to 1366 x 768, and most systems reportedly run on a 14"–15.6" monitor. The font size is 50pt, which is equivalent to approximately 6mm x-height. Distance between baselines was approximately 18mm with the most commonly used font–otherwise 24mm.



(a) Words per cleaned reading session

(b) Misreading ratio per cleaned session

Figure 2: Distributions of total number of words and misreading ratios per session after cleaning.

3.1.2 Subjects

The cleaned dataset contains 44 unique readers with different reading durations. Readers are probably between 5 and 15 years old, which is the official age of students in the Danish schools, but we do not know their exact ages. To control for reading proficiency, we include the texts' readability scores as a feature in all experiments. All students receive extra reading classes, because they struggle with reading. Many of them are probably dyslexic, but we do not have access to this information. Because this is not experimental data, the students will have received different instructions from the teachers. We do not know if they picked the text themselves or for how long they read prior to each recording. They are not necessarily alone in the room, but it is a fair assumption that they all make an effort to read correctly because they are recorded. The data comes from a number of different systems that we were informed is in the range between 10 and 20, but the actual number of schools and teachers is unknown to us. All children and their parents gave consent that the anonymized eye-tracking data may be used for this research.

3.2 Features

Reading patterns have been shown to be influenced by a number of factors, including textual features and the instructions given to a reader, such as encouraging a specific reading strategy. Readers, or different groups of readers, furthermore display individual reading styles which affect the eye movements (Benfatto et al., 2016). Other factors include the reader's individual skill level, cognitive abilities and mood, among others.

We extract a number of gaze features that have been associated with processing load. Some of our gaze features directly reflect the processing load associated with a word, especially the two correlated measures *total fixation duration* and *number of re-fixations*, but also the *mean fixation duration*. Some gaze features are included to account for preview effects (whether the next or previous word was fixated) as well as the scan path immediately surrounding the word. We split the gaze features into two groups: GAZE (W) for features directly associated with word-level processing and GAZE (C) for features associated with the eye movements on the immediate context of the word. All features are scaled to the $[-1, 1]$ interval.

We further extract a number of basic features that are known to affect gaze features and thus need to be controlled for. These include word length and word frequency (Hyönä and Olson, 1995), but also position in sentence (Rayner et al., 2000) and position on the page have shown to affect reading for adults. We also include a range of linguistic features that we expect to describe word difficulty. All features and feature groups are listed in Table 2 and described below.

Gaze features During reading, the reader performs a series of stable fixations of a couple of hundred milliseconds duration on average. Between fixations, the eyes perform rapid, targeted movements, called *saccades*. All gaze features are computed on the word level and use the application's definition of the area of interest surrounding each word.

For gaze duration, we extract both late and early processing measures. Late measure such as total *fixation duration* and *number of re-fixations* reflect late syntactic and semantic processing in skilled adult reading (Rayner et al., 1989). For children with reading difficulties, we assume these measures to likely reflect processing difficulty.

For the first three passes over a word, we also extract the direction and the word distance of both the ingoing and outgoing saccade.[4] These six features are expected to map the activity around the word and, for example, show whether some word was part of sequential, forward reading or occurred in a series of erratic saccades.

Four features indicate the *landing positions* of fixations in four equally-sized parts of the display width of a word. This captures whether a word, for instance, has three fixations on the last quarter of its display width, which would be atypical and suggest that the reader is struggling with the ending of this word. We further explicitly encode the landing position of the first and last fixation. Note that because of the anatomy of the eye, eye tracking can never be pixel-accurate, but has at least 2° inaccuracy. For short words (or words printed very small, which does not apply for this study) these features may be misleading.

The data also provides pupil sizes for both eyes. It is well known that the pupil dilates as response to external lighting factors, but there is also evidence that the pupil systematically–but on a much smaller scale–dilates as a response to mental state, emotions or concentration (Beatty et al., 2000). In an experiment collecting pupil size, one would control lighting, which was not possible in the present scenario. For all pupil measures, we subtracted the same side mean of the reading session. We confirmed that all changes larger than 0.6 times the mean were captured when removing short fixations, as they may be caused by the tracker mistaking eyelashes for pupils during blinks.

Basic features The basic features span 16 textual and presentational features that are either directly accessible via the system or easily obtainable. They are included in all our experiments and serve as control features for the gaze features because we expect them to explain some of the variance in the gaze features, e.g. reading changes

[4] As we removed everything that was not a fixation on text before calculating the gaze features, intermediary non-text fixations may have occurred between text fixations, such as image fixations. We count the last/next fixated *word*. For example, if a word has index 5, and the first pass incoming saccade is from word index 4, we get a feature value of -1 for first pass ingoing.

BASIC	GAZE ON WORD (W)
Is bold	Number of fixations on word
Is italic	First fixation duration
Is lowercase	Mean fixation duration
Is uppercase	Total fixation duration
Has punctuation	Count of passes over the word
Line index on page	Left pupil size
Word index on line	Right pupil size
Page number	Refixation counts
Position in sentence (relative)	Fixations in first quarter count
Position in sentence (absolute)	Fixations in second quarter count
Sentence length (characters)	Fixations in third quarter count
Sentence length (words)	Fixations in fourth quarter count
Word index	Relative landing position of first fixation
Sentence index	Relative landing position of last fixation
Word length (characters)	Average character index of fixations
GAZE IN CONTEXT (C)	LINGUISTIC
$1st$ pass ingoing saccade dist. and dir.	LIX score for entire text
$1st$ pass outgoing saccade dist. and dir.	Previous occurrences of word stem in text
$2nd$ pass ingoing saccade dist. and dir.	Previous occurrences of word type in text
$2nd$ pass outgoing saccade dist. and dir.	Vowel count
$3rd$ pass ingoing saccade dist. and dir.	Character perplexity
$3rd$ pass outgoing saccade dist. and dir.	Word frequency
Next word fixated	Universal POS tag
Previous word fixated	

Table 2: Overview of the feature groups used in the experiments.

over the course of a line and the course of a sentence (Just and Carpenter, 1980). We further encode the line number a word is located in on a page, as well as its position in that line.

Linguistic features The linguistic features include the absolute vowel count, which in Danish is highly correlated with the number of syllables. Universal POS tags are obtained from the Danish Polyglot tagger.[5] We also include the provided *Läsbarhetsindex* (LIX) (Björnsson, 1968), a Swedish readability metric (commonly also applied to Danish) that considers the mean sentence length and the ratio of long words (more than 6 characters). The log word probability is estimated from a language model we train on the entire Danish Wikipedia (downloaded in November 2017) using KenLM (Heafield, 2011). Frequency affects processing load and thus fixation duration for adults as well as dyslexic and neurotypical Finnish children (Hyönä and Olson, 1995), but there is conflicting evidence whether text frequencies from adult text explain variance in children's eye movements (Blythe and Joseph, 2011). Character perplexity is estimated using a 5-gram character language model, also using KenLM on the Danish Wikipedia. The previous occurrence of stems and word types is included as reading time for low-frequency words has shown to decrease on later repeats in a text (Rayner et al., 1995). We use NLTK's snowball stemmer for Danish.

4 Model

In preliminary experiments, we observed that the relatively small overall amount of data, as well as the low fraction of positive instances, caused significant variation between repeated random

[5] `http://polyglot.readthedocs.io`

Feature group	F_1	
BASIC	18.78	†
+ GAZE (W)	40.50	*
+ GAZE (C)	18.49	†
+ LINGUISTIC	19.24	†
+ GAZE (W) + GAZE (C)	**41.19**	*
+ GAZE (W) + LINGUISTIC	41.08	*
+ GAZE (W) + LINGUISTIC	18.65	†
All features	40.42	*

Table 3: Performance across feature groups for Experiment 1. Scores are averaged F_1 over ten cross-validation folds. Using an independent t-test, * and † indicate results from ten cross validation rounds significantly different from BASIC and the best feature combination BASIC + GAZE(W) + GAZE(C), respectively.

restarts of various classification algorithms. We thus approach the task of predicting misreadings from gaze with ensemble methods, training N classifiers independently on the same data and letting them vote on the instances in a held-out development set. Using this development set, we then optimize a threshold t, which is the fraction of the number of classifiers that need to cast a positive vote on an item before we accept it as such.

All of our ensembles consist of 10 random forest classifiers and 10 feed-forward neural networks. The random forests, in turn, consist of 100 trees that create splits based on Gini impurity (Breiman, 2001). The neural network models are implemented in Pytorch and trained with the Adam algorithm (Kingma and Ba, 2014), with an initial learning rate of $3 \cdot 10^{-4}$ and a dropout rate of 0.2 on the hidden layers, whose number and sizes we vary in our experiments. We further employ early stopping, monitoring the loss on the development set with a patience of 30 steps.

4.1 Multi-task learning for cross-user knowledge transfer

One of the central questions we investigate in this paper is to what degree gaze patterns for misread words vary between readers, and whether we can learn to transfer knowledge about predictors of misreadings between readers. We address these questions in the experiments reported in Section 5.2, for which we use a multi-task learning (MTL) model that employs hard parameter sharing. MTL has received significant attention in the natural language processing community over the past years (see Bjerva (2017) for a review). One of the most intriguing properties of MTL is that it allows for the transfer of knowledge between different tasks and datasets, which has been investigated and exploited in a growing number of works (Klerke et al., 2016; Martínez Alonso and Plank, 2017; Bingel and Søgaard, 2017), including work on the identification of complex words (Bingel and Bjerva, 2018).

In this work, we view the different readers as different *tasks*, motivated by Bingel and Bjerva (2018), who interpret different languages as different tasks for cross-lingual complex word identification. We define a feed-forward neural network model with one output layer per reader, all of which are dense projections from a shared hidden layer. In this framework, each training step consists of flipping a coin to sample any of the tasks and retrieving a batch of training data for this task. This batch is then used to optimize both the shared and the respective task-specific parameters. For a detailed definition of the model, see Bingel and Bjerva (2018).

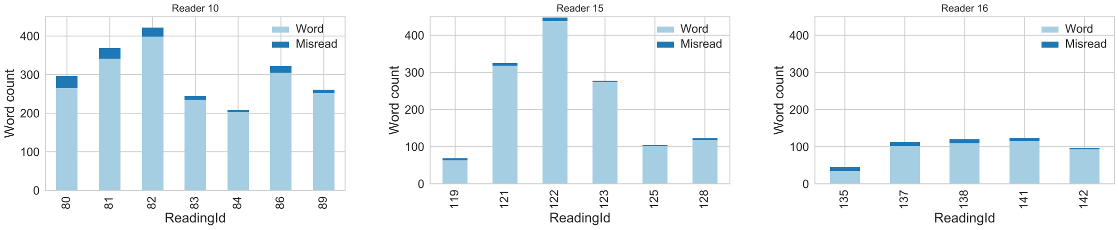
5 Experiments

5.1 Experiment 1: Across entire dataset

As a first experiment, we investigate the performance of our models and the predictiveness of the individual feature groups through 10-fold cross validation across the entire dataset. At each fold, we reserve one tenth of the data for testing and another tenth to monitor validation loss of the network as the early stopping criterion.

Note that we split the data randomly and do not stratify the cross-validation splits in any way. In conjunction with the strong class imbalance, this means that we are likely to encounter very different class distributions across splits. This setup may generally lead to lower performance scores, likely with greater variance. However, this was a deliberate choice as we cannot assume a consistent class distribution across train and test set in the real world, or in fact hardly any prior knowledge with regards to class distribution in the test set. Random splitting also means that data from the same *reading* will likely be distributed across train and test partitions for a certain cross-validation iteration.

We perform a first baseline experiment with



(a) Reader 10

(b) Reader 15

(c) Reader 16

Figure 3: Words and misreading counts for readings of three readers in cross-user experiment

UserId	Number of	Words per reading		Thereof misread	
	reading sessions	Mean	std.dev.	Mean	std.dev.
10	7	285.9	67.5	16.6	9.9
15	6	219.2	148.1	5.0	2.3
16	5	91.6	32.7	8.0	3.1

Table 4: Statistics of (misread) words in sessions for the three readers with most readings.

only the basic features that we list in Section 2. On top of this baseline feature set, we perform further experiments, incorporating all combinations over the other feature groups. The results we present in Table 3 are based on the best respective model architecture for each feature combination, evaluated via the average over validation splits.[6]

5.2 Experiment 2: Cross-reader prediction

Without reader's own data In a second experiment, we are interested in how well our model can predict misreadings for specific readers. For this, we identify the three readers with most reading sessions and perform a range of experiments, testing our models on the readings of each of these readers after training them on all other data. We denote the three most active readers by their unique, anonymized IDs as they appear in the dataset: 10, 15 and 16. These readers have 7, 6 and 5 recorded and marked readings, respectively, and we present statistics on these readings in Table 4 and Figure 3. As in the previous experiment, we optimize our model through cross validation to tune hyperparameters and perform early stopping. We report test data results for the model with optimal validation performance in Figure 4, broken down into each reader's different sessions.

Learning from reader's own data Complementing the setup above, we now investigate how data from the same reader, but from different reading sessions, can inform our models. Therefore, we further perform cross-validation experiments across each reader's sessions. More concretely, for a reader with n marked readings, we perform n-fold cross validation, holding out one reading a time as a test set and another to monitor validation loss for early stopping of the neural model, while training on the remaining $n - 2$ readings.

MTL As outlined in Section 4.1, we now view readers as tasks in an MTL model. For each of the three readers identified above and for each test reading, we train an ensemble whose neural MTL models define two outputs: one for the reader in question and one combined output for all other readers in the entire dataset. The random forest classifiers are trained on all remaining data except the held-out validation and test readings.

6 Results and Discussion

From Experiment 1, we observe that gaze features of the target word itself contribute strongly to model improvements over the baseline of textual features (see Table 3). Contextual gaze features and linguistic features do so to a lesser degree. The best feature group combination consists of the basic features and both gaze feature groups. Adding the linguistic features to this seems to slightly dilute the model.

[6] To address the variation in input dimensionality as we consider different feature group combinations, we train models with different architectures: (i) a single hidden layer with 20 units, (ii) two hidden layers with 20 units each, and (iii) a single hidden layer with 40 units.

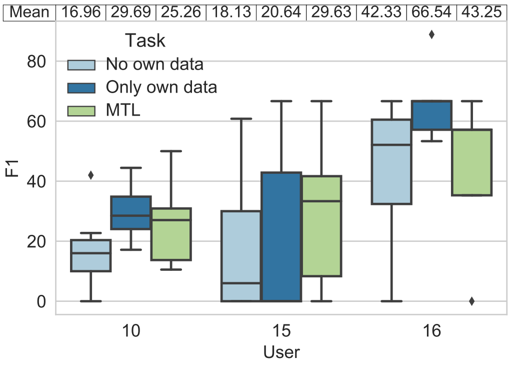


Figure 4: F_1 score distributions across test readings for each of the three readers with most sessions for three tasks.

The results from Experiment 2 in Figure 4 show that, at least for these three readers, there is a considerable degree of specificity attested in the reading patterns of misread words: in the scenario where we learn only from other users' gaze patterns (shown in light blue), performance is generally worse than for the other approaches. The high degree of reader specificity is also reflected in the comparison between learning just across a single user's readings and a multi-task setup that also considers other readers. Here, we observe that the former attains higher mean F_1 scores across readings for readers 10 and 16, although MTL is superior to the single-task setup for reader 15. Another observation is that misreadings can generally be predicted much better for reader 16 than for the other readers, which may in part be due to the higher ratio of misread words in these readings.

As especially our cross-reader experiments show, there is reason to believe that the manifestations of misreadings in gaze differ strongly between these readers. However, since we do not have information on the individual readers' age or general reading proficiency, we cannot confidently conclude whether the better stability of within-user experiments attested in Figure 4 is due to reader-specific idiosyncrasies or group-internal patterns (which would be supported by evidence that readers 10 and 16 were more atypical readers than others in the present dataset). We find some support for the latter hypothesis in literature describing children's reading development, which identifies a range of patterns common to young and low-proficiency readers. These patterns include longer and more frequent fixations, shorter saccadic amplitude and more regressions – all of which are also associated with comprehension difficulties, see Blythe and Joseph (2011) for a review. The presence of group-internal patterns is further supported by the observation that we are still able to successfully transfer knowledge about readings patterns between users in some cases, increasing performance for the readings of user 15.

One disadvantage of noisy, real-world data is that we do not know to what degree similarities and differences in the data, as well as our results, are influenced by chance, or whether they will generalize to other gaze data. The fact that many parameters are outside of our control and also outside of our knowledge means that we cannot describe certain biases in the data (such as age or reading skill) and consider them as causes for statistical variations in model performance.

7 Conclusion

This paper presented first work in the automatic prediction of reading errors in children with dyslexia and other reading difficulties using real-world gaze data. We showed that despite the noisy conditions under which this data was obtained, features we extract from the gaze patterns are predictive of reading mistakes among children. Besides the immediate application in automating some parts of reading teaching, this could be exploited in personalized text simplification, where gaze could be used as feedback to the system.

Our experiments further show that while gaze patterns for misreadings seem to be largely specific to individual readers or groups of readers, we can successfully use MTL to transfer knowledge between readers at least in some cases. Note also that we have very little knowledge of the age and general proficiency of specific readers, including those investigated in our MTL experiments, and we expect that our MTL approach can be much more successful between more similar readers.

Acknowledgements

The authors wish to thank the children using EyeJustRead, as well as their parents, for giving us permission to use their data. We are also grateful to Emil Juul Jacobsen and Janus Askø Madsen for supplying us with the data. We further acknowledge valuable comments by Anders Søgaard as well as by the anonymous reviewers. We acknowledge the support by Trygfonden.

References

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*, pages 579–584.

Jackson Beatty, Brennis Lucero-Wagoner, et al. 2000. The pupillary system. *Handbook of psychophysiology*, 2:142–162.

Mattias Nilsson Benfatto, Gustaf Öqvist Seimyr, Jan Ygge, Tony Pansell, Agneta Rydberg, and Christer Jacobson. 2016. Screening for dyslexia using eye tracking during reading. *PloS one*, 11(12):e0165508.

Joachim Bingel and Johannes Bjerva. 2018. Cross-lingual complex word identification with multitask learning. In *Proceedings of the Complex Word Identification Shared Task at the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, New Orleans, United States. Association for Computational Linguistics.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. *15th Conference of the European Chapter of the Association for Computational Linguistics*.

Johannes Bjerva. 2017. One model to rule them all: Multitask and multilingual modelling for lexical analysis. *arXiv preprint arXiv:1711.01100*.

Carl Hugo Björnsson. 1968. *Läsbarhet*. Liber.

Hazel I Blythe and Holly SSL Joseph. 2011. Children's eye movements during reading. *The Oxford Handbook of Eye Movements*, pages 643–662.

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

Ana Valeria Gonzalez-Garduño and Anders Søgaard. 2017. Using gaze to predict text readability. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 438–443.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

Jukka Hyönä and Richard K Olson. 1995. Eye fixation patterns among dyslexic and normal readers: Effects of word length and word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(6):1430–40.

Interagency Committee on Learning Disabilities. 1987. Learning Disabilities: A Report to the U.S. Congress. Technical report, Government Printing Office, Washington DC, U.S.

Holly SSL Joseph, Simon P Liversedge, Hazel I Blythe, Sarah J White, Susan E Gathercole, and Keith Rayner. 2008. Children's and adults' processing of anomaly and implausibility during reading: Evidence from eye movements. *Quarterly Journal of Experimental Psychology*, 61(5):708–723.

Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological review*, 87(4):329–354.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Sigrid Klerke, Héctor Martínez Alonso, and Anders Søgaard. 2015a. Looking hard: Eye tracking for detecting grammaticality of automatically compressed sentences. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 97–105.

Sigrid Klerke, Sheila Castilho, Maria Barrett, and Anders Søgaard. 2015b. Reading metrics for estimating task efficiency with MT output. In *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*, pages 6–13.

Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. 2016. Improving sentence compression by learning to predict gaze. *NAACL*, pages 1528—-1533.

Sigrid Klerke, Janus Askø Madsen, Emil Juul Jacobsen, and John Paulin Hansen. 2018. Substantiating reading teachers with scanpaths.

Héctor Martínez Alonso and Barbara Plank. 2017. When is multitask learning effective? Semantic sequence prediction under varying data conditions. In *15th Conference of the European Chapter of the Association for Computational Linguistics*.

Marcus Nyström and Kenneth Holmqvist. 2010. An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior research methods*, 42(1):188–204.

Gustavo Paetzold and Lucia Specia. 2016. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

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

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.

Keith Rayner, Gary E Raney, and Alexander Pollatsek. 1995. Eye movements and discourse processing. pages 241—-255.

Keith Rayner, Sara C Sereno, Robin K Morris, A Rene Schmauder, and Charles Clifton Jr. 1989. Eye movements and on-line language comprehension processes. *Language and Cognitive Processes*, 4(3-4):SI21–SI49.

Matthew Shardlow. 2014. Out in the open: Finding and categorising errors in the lexical simplification pipeline. In *LREC*, pages 1583–1590.

Matthew J Traxler. 2002. Plausibility and subcategorization preference in children's processing of temporarily ambiguous sentences: Evidence from self-paced reading. *The Quarterly Journal of Experimental Psychology: Section A*, 55(1):75–96.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, New Orleans, United States. Association for Computational Linguistics.