A Survey on Using Gaze Behaviour for Natural Language Processing

Sandeep Mathias[1] , **Diptesh Kanojia**[1,2] , **Abhijit Mishra**[1*] and **Pushpak Bhattacharyya**[1]

[1]Indian Institute of Technology Bombay
[2]IITB-Monash Research Academy
{sam, diptesh, pb}@cse.iitb.ac.in, abhijitmishra.530@gmail.com

Abstract

Gaze behaviour has been used as a way to gather cognitive information for a number of years. In this paper, we discuss the use of gaze behaviour in solving different tasks in natural language processing (NLP) without having to record it at test time. This is because the collection of gaze behaviour is a costly task, both in terms of time and money. Hence, in this paper, we focus on research done to alleviate the need for recording gaze behaviour at run time. We also mention different eye tracking corpora in multiple languages, which are currently available and can be used in natural language processing. We conclude our paper by discussing applications in a domain - education - and how learning gaze behaviour can help in solving the tasks of complex word identification and automatic essay grading.

1 Introduction

Collecting psycholinguistic information from a reader has benefited multiple tasks in NLP, like named-entity recognition (NER) [Hollenstein and Zhang, 2019], text quality rating prediction [Mathias *et al.*, 2018], sarcasm understandability [Mishra *et al.*, 2016a], *etc*. Gaze behaviour, in particular, has been shown to correlate well with cognitive processing of text, via the eye-mind hypothesis, which states that "there is no appreciable lag between what is fixated by the eye and what is processed by the mind" [Just and Carpenter, 1980]. Therefore, while using gaze behaviour is helpful for solving NLP tasks, a massive challenge involved is *how do we collect the gaze behaviour in the first place*? In this paper, we describe research that uses gaze behaviour at run time to solve different NLP tasks.

While studying the gaze behaviour of a reader, we define the following terms. An **Interest Area** is the part of the screen that is of interest for analysis. In the case of NLP, it is mainly the words, although it could also be phrases, sentences, and paragraphs. A **Fixation** is an event where the eye is focused on the screen. Fixations take place when the eye is processing what it sees - in this case, the text that is being

read. A **Saccade** is the movement of the eye from one fixation point to the next. There are 2 types of saccades - **Progressions** and **Regressions**. Progressions take place when there is a saccade from a fixation in the current interest area to a fixation in a later interest area. Regressions take place when there is a saccade from a fixation in the current interest area to a fixation in an earlier interest area. Figure 1 shows the output of an eye-tracker (the SR Research Eye Link 1000), for a reader reading a piece of text. The yellow boxes correspond to the interest areas, the cyan circles correspond to fixations, and the dark-yellow arrows correspond to saccades.

2 Motivation for Learning Gaze Behaviour

One of the earliest works that postulated the utility of gaze behaviour was a study on reading comprehension done by Just and Carpenter [1980]. They came up with the eye-mind hypothesis, which stated that "there is no appreciable lag between what is fixated by the eye and what is processed by the mind." Many studies, done in the field of cognitive science, have found relationships between different aspects of gaze behaviour and corresponding aspects of text, such as relationships between fixations and word length [Rayner, 1998; Henderson and Ferreira, 1993], word predictability [Rayner, 1998], etc.

The gaze behaviour of a reader can therefore provide valuable psycholinguistic information for systems to help them solve tasks that require human intelligence. For example, using gaze behaviour, we can better quantify the difficulty of sentences for translation [Mishra *et al.*, 2013], or verify if a reader understands the sarcasm in a piece of text [Mishra *et al.*, 2016a], or even evaluate the quality of word embeddings [Søgaard, 2016]. Portable devices and applications (like Samsung Smart Scroll) generally provide a lower quality of gaze behaviour recording compared to research-grade devices. Our goal in this paper is to introduce the AI audience to ways in which we alleviate the need for readers to read texts at run time for the purpose of collecting gaze behaviour.

3 Eye Tracking Corpora

There are several publicly available corpora where readers solve different tasks in different languages[1]. One of the earli-

---

*currently employed at Apple Inc.

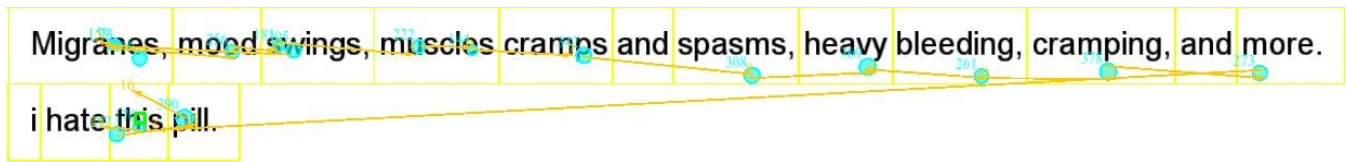[1]A number of those corpora can be downloaded from here.

Figure 1: Definitions of different terms used for gaze behaviour. The yellow boxes represent interest areas, the cyan circles represent fixations, and the dark yellow arrows correspond to saccades. Source: [Mishra *et al.*, 2016c].

est publicly available eye-tracking corpus is the Dundee Corpus [Kennedy *et al.*, 2003]. The Dundee Corpus was created for both English and French, with the English version having 20 articles from *The Independent* read by 10 readers, and the French version having 20 articles from *Le Monde* read by 10 French speakers.

There are a lot of corpora available in English. Some of them merely capture a reader's gaze behaviour as they read texts varying from articles to a novel. Others capture the reader's gaze movement as the readers solve different NLP tasks. Along with English and French, there are a number of other publicly available corpora in other languages, like Chinese [Zang *et al.*, 2018; Li *et al.*, 2018], Dutch [Cop *et al.*, 2017; Mak and Willems, 2019], German [Nicenboim *et al.*, 2016; Kliegl *et al.*, 2004] Persian [Safavi *et al.*, 2016], Russian [Laurinavichyute *et al.*, 2017], and Spanish [Nicenboim *et al.*, 2016]. Table 1 gives the statistics for each of the publicly available eye-tracking corpora.

4 NLP Tasks Where Gaze Behaviour is Used

Gaze behaviour is also beneficial for multiple tasks in NLP where the reader's psychological input is critical, and text features alone will not be enough. Mishra *et al.* [2013] discuss how using gaze behaviour would be a better approach to judge the complexity of translating a sentence, compared to using just length-based statistics (like word length, sentence length, *etc.*). Understanding sarcastic texts could also be resolved using gaze behaviour, where incongruity in the text (one of the leading indicators of sarcasm) induces gaze behaviour characterized by longer fixations, regressions *etc.* [Mishra *et al.*, 2016a]. Gaze behaviour has also been used to identify a reader's native language [Berzak *et al.*, 2017], as well as in detecting grammatical errors in compressed sentences [Klerke *et al.*, 2015b]. Klerke *et al.* [2015a] also show that gaze behaviour can be used to evaluate the output of Machine Translation systems better than automated metrics. Gaze behaviour has also been used to evaluate how a reader would rate the quality of a piece of text [Mathias *et al.*, 2018].

The scanpath of a reader (*i.e.* the path a reader's eye traverses when they read the text) has been used to test how easy/difficult a piece of text is for a reader to read [Mishra *et al.*, 2017]. It can also be used to predict the misreadings among children with reading difficulties [Bingel *et al.*, 2018].

Gaze behaviour has also been used in multiple areas of sentiment analysis such as sarcasm detection [Mishra *et al.*, 2016b], sarcasm understanding [Mishra *et al.*, 2016a], and sentiment analysis annotation tasks [Joshi *et al.*, 2014]. Klerke and Plank [2019] tackle the classical problem of PoS

(Part-of-speech) tagging with gaze data and provide a systematic overview of the influence of two independent levels of gaze data aggregation on low-level syntactic labelling tasks at two separate levels of complexity; *i.e.,* a simple chunk boundary tagging and a supervised PoS tagging task.

5 Learning Gaze Behaviour

As mentioned in the previous section, using gaze behaviour helps systems in solving many NLP tasks. However, using gaze behaviour, while alleviating the need to record it at test / run time, is still a relatively new challenge. In this section, we look at different tasks and different systems which alleviate the need for collecting gaze behaviour at run time.

5.1 NLP Tasks

Predicting Fixations while Reading. Nilsson and Nivre [2009] describe an approach to detect which tokens readers fixate on while reading, using a transition-based approach. They use a transition-based model of reader's eye movements to predict the next word that is fixated. They used features like token length, token frequency class, next token length, next token frequency class, etc. Matthies and Søgaard [2013] improve on their approach using a linear CRF model to determine which words are fixated on while reading. The features that they use are word length (for a window of 5 words), and word probability (for a window of 3 words). While fixation accuracy and F1 are comparable to Nilsson and Nivre [2009]'s approach, Matthies and Søgaard [2013]'s approach predicts the eye movements better for new readers, than when training and test data come from the same reader.

Predicting Grammatical Functions. Barrett and Søgaard [2015b] describe an approach to predict the grammatical functions of words in a sentence (i.e. whether they are subject, object, etc.). They trained their system using gaze behaviour data from Barrett and Søgaard [2015a], and found statistically significant improvements over their baseline, using a L2-regularized logistic regression classifier.

Text Simplification. Klerke *et al.* [2016] describe a way to simplify text by compressing sentences using gaze behaviour learnt at run time. They used the Ziff-Davis [Knight and Marcu, 2002], Broadcast [Clarke and Lapata, 2006], and Google [Filippova *et al.*, 2015] datasets and multi-task learning, where the primary task is compressing the sentences, and the auxiliary tasks are learning the gaze behaviour attributes, namely the first fixation duration, and regression duration. They describe two multi-task learning models for using gaze behaviour. The first approach (called Multi-task in their paper) uses multi-task learning with a separate logistic

Dataset	Source	Language	Stimulus	Total Subjects
Zang et al. (2018)	[Zang *et al.*, 2018]	Chinese	90 sentences	35
Li et al. (2018)	[Li *et al.*, 2018]		15 documents	29
GECO	[Cop *et al.*, 2017]	Dutch	1 novel	33
Mak & Willems (2019)	[Mak and Willems, 2019]		3 short stories	102
ZuCo	[Hollenstein *et al.*, 2018]	English	1107 sentences	12
GECO	[Cop *et al.*, 2017]		1 novel	33
Parker et. al (2017)	[Parker *et al.*, 2017]		40 documents	48
PROVO	[Luke and Christianson, 2018]		55 texts	84
ASD Data	[Yaneva, 2016]		27 documents	27
CFILT-Quality	[Mathias *et al.*, 2018]		30 documents	20
CFILT-Scanpath	[Mishra *et al.*, 2017]		32 documents	16
CFILT-Coreference	[Cheri *et al.*, 2016]		22 documents	14
CFILT-Sarcasm	[Mishra *et al.*, 2016a]		1000 tweets	7
CFILT-Sentiment	[Joshi *et al.*, 2014]		1059 sentences	5
UCL Corpus	[Frank *et al.*, 2013]		205 sentences	43
Dundee	[Kennedy *et al.*, 2003]		20 documents	10
Dundee	[Kennedy *et al.*, 2003]	French	20 documents	10
Self-Paced Reading Time	[Nicenboim *et al.*, 2016]	German	176 sentences	72
Potsdam Sentence Corpus	[Kliegl *et al.*, 2004]		144 sentences	55
Dependency Resolution dataset	[Safavi *et al.*, 2016]	Persian	136 sentences	40
Russian Sentence Corpus	[Laurinavichyute *et al.*, 2017]	Russian	144 sentences	96
Self-Paced Reading Time	[Nicenboim *et al.*, 2016]	Spanish	212 sentences	79

Table 1: Summary of eye-tracking datasets available today. The novel read in the GECO dataset is *The Mysterious Affair at Styles* by Agatha Christie (in both Dutch and English). Dataset is the name of the dataset. Source is the reference where the dataset was published. Language is the language of the data in the dataset. Stimulus is the amount of text data that the annotators read. Total subjects is the number of subjects who participated in the creation of the corpus.

regression classifier for learning gaze behaviour during training. The second approach (Cascaded) also uses multi-task learning with gaze behaviour learnt from an inner layer in their network.

Part-of-Speech (PoS) tagging. Barrett *et al.* [2016a] describe an approach to solve the task of part-of-speech tagging using gaze behaviour from the Dundee Treebank [Barrett *et al.*, 2015]. Barrett *et al.* [2016b] investigate the same using cross-lingual approaches (i.e. training on English, test on French, and vice versa). Both approaches use a Hidden Markov Model with additional type-aggregated gaze features. The features that they used were *early* gaze features (which measure events when the reader first encounters a word, like first fixation duration), *late* gaze features (which measure events when the reader finishes reading the word, like number of regressions to a word), *basic* gaze features (measures which don't specifically belong to either the early or late gaze features, like dwell time), Regression-from features (which capture regressions from a word, like total regression_from duration), context gaze features (namely the fixation durations and probabilities of neighbouring words), and finally, the NoGaze features, which are text features from the British National Corpus and Dundee Corpus [Kennedy *et al.*, 2003].

Readability. González-Garduno and Søgaard [2018] describe a solution for predicting readability using multi-task learning, with readability prediction as the primary task, and predicting gaze behaviour as an auxiliary task. They use a multi-task multi-layer perceptron and multi-task logistic regression systems, with their best results achieved by using all

features with the multi-task multi-layer perceptron, with their gaze behaviour learnt from the Dundee Corpus.

Sentiment Analysis. Mishra *et al.* [2018] describe a way to use gaze behaviour for the task of sentiment analysis of movie reviews. They use a multi-task framework, where the primary task is predicting the sentiment, and the auxiliary tasks are PoS tagging and learning gaze behaviour. They use a pair of bi-directional LSTM layers, where each layer performs a particular task (either PoS tagging or learning gaze behaviour). They showed statistically significant improvements over the state-of-the-art reported results using both their multi-task systems on the IMDB25K dataset [Maas *et al.*, 2011], and the PL2000 dataset [Pang and Lee, 2004].

Named Entity Recognition (NER). Hollenstein and Zhang [2019] describe an approach to perform named entity recognition, using gaze behaviour aggregated over their gaze corpora showing improvements on the CoNLL 2003 dataset [Tjong Kim Sang and De Meulder, 2003]. To train their system to use gaze behaviour, they used the Dundee Corpus [Kennedy *et al.*, 2003], GECO Corpus [Cop *et al.*, 2017], and ZuCo Corpus [Hollenstein *et al.*, 2018] for collecting the gaze data. For each token, they extract gaze features from the gaze corpora using type aggregation. Similar to Barrett *et al.* [2016a], Hollenstein and Zhang [2019] use similar gaze features - namely early, late, basic and context gaze features. For each word, their system takes character and word embeddings, along with the gaze features as input, and returns the corresponding NER tag. To do this, they use a bi-directional LSTM neural network with conditional random fields to per-

form NER.

Sequence Classification. Barrett *et al.* [2018] describe a solution to multiple NLP tasks - sentiment analysis, grammatical error detection, and hate speech detection - using multitask learning. They use a bi-directional LSTM [Hochreiter and Schmidhuber, 1997] layer with attention for encoding each token to get a sentence representation. The network's attention weights are learnt using existing gaze behaviour data, so that no gaze data recording is done at run time. They show that biasing the network's attention using gaze behaviour leads to performance improvements across all 3 tasks as well as aid in explaining the network's decisions.

5.2 System Architectures

One of the earliest frameworks to predict gaze behaviour was EZ-Reader [Reichle *et al.*, 2003], which was used for understanding how word identification, visual processing, attention, and oculomotor control jointly determine when and where the eyes move during reading. Another system, called Swift [Engbert *et al.*, 2005] used a mathematical model for the control of eye movements during reading that is both psychologically and neurophysiologically plausible and that accounts for most of the then known experimental findings.

There are other different approaches used today to alleviate the need for recording gaze behaviour at run time. Feature engineering is a way in which we predict the gaze behaviour data using lexical properties, such as word length, syllable count, word surprisal, etc. Type aggregation is another way in which we avoid tracking a reader's eyes when they read the text. In this approach, we use a gaze corpus (like the Dundee Corpus) and pre-train our model with aggregate values of each gaze feature for each word. Multi-task learning is an approach where learning gaze behaviour is an auxiliary task. This approach is often used to learn gaze at the word-level, but our task requires an output at the sentence / document level.

5.3 Normalizing Data

While collecting gaze behaviour data, it is better to normalize gaze behaviour across readers. This is done because readers may read the text either faster, or slower, depending on their fluency with the language. Therefore, we must normalize the gaze behaviour data before use. Gaze behaviour is normalized in the following ways. **Min-Max Normalization** is where we normalize the data on a scale of $[0, 1]$, where 0 corresponds to the lowest value, and 1 corresponds to the highest value. Barrett *et al.* [2016a] is one work that uses min-max normalization. **Binning** is where we consider discrete bins for each gaze behaviour attribute. For example: 6 bins (numbered 0 to 5) as described by Klerke *et al.* [2016]. One of the advantages of binning over min-max normalization is that it reduces the effect of outliers in the data. In both approaches of normalization, the normalization takes place for each reader.

6 Applications of Learning Gaze Behaviour

In this section, we cover some applications in NLP which we believe could benefit a lot from *learning* gaze behaviour. To the best of our knowledge, there is no work on any of these applications where the system learns gaze behaviour.

6.1 Complex Word Identification

Lexical simplification is a process in which text gets simplified by replacing complex words and phrases with simpler ones. For example, a non-native speaker of English will struggle to understand what the word "procrastinate" means, in the sentence "We should not procrastinate our submission." However, they will be more likely to understand the meaning of the sentence "We should not delay our submission." The process of identifying which words are hard (and should be replaced by an appropriate synonym) is a very useful application for the use of eye-tracking, as readers fixate longer on harder words compared to easier words [Rayner, 1998].

Paetzold and Specia [2016] report the results of the shared task on complex word identification held at Sem-Eval 2016. Another shared task was organized in 2018, to identify complex words, and phrases, in English, Spanish, German, and a cross-lingual setting (where the target language was French) [Yimam *et al.*, 2018].

Kunze *et al.* [2013] showed how eye tracking can be used for identifying complex words. However, their approach required readers to read the text, while they track their eye movements, in order to identify complex words. Therefore, using cognitive information should aid in solving these tasks. As mentioned earlier, there are quite a few papers that deal with text complexity / simplification and readability which use gaze behaviour like Klerke *et al.* [2016], Mishra *et al.* [2017], González-Garduño and Søgaard [2017], González-Garduno and Søgaard [2018], etc. However, in proposing solutions for the shared tasks, none of them used cognitive information. With the availability of a large number of eye-tracking corpora, in various languages, an interesting avenue of research would be exploring how using gaze behaviour can help in identifying complex words, even if we don't have such information at run time.

6.2 Automatic Essay Grading

An essay is a text, written in response to a topic, called the essay prompt. Grading an essay is assigning a score to the essay based on its quality, either for the essay on the whole (holistic scoring) or for certain aspects of the essay (trait-specific scoring) [Ke and Ng, 2019]. Automatic essay grading (AEG) is the process of grading an essay using a machine. The earliest AEG system was described by Page [1966], over 50 years ago. Since then, there have been a number of commercial AEG systems, like E-Rater [Attali and Burstein, 2006], Intelligent Essay Assessor [Landauer, 2003], LightSide [Mayfield and Rosé, 2013], *etc.*

The current state-of-the-art AEG systems use neural networks, like CNNs [Dong and Zhang, 2016], LSTMs [Alikaniotis *et al.*, 2016; Taghipour and Ng, 2016; Tay *et al.*, 2018], or both [Dong *et al.*, 2017; Zhang and Litman, 2018; Liang *et al.*, 2018]. The dataset that they use is the 2012 Automatic Student Assessment Prize (ASAP) AEG Dataset, released by the Hewlett Foundation [Ke and Ng, 2019].

As mentioned earlier, Mathias *et al.* [2018] describe a way to predict the rating a reader would give a piece of text based

on its quality. Their work showed that gaze behaviour could be used for similar applications like AEG. However, their approach required readers to read the text in order to use the reader's gaze behaviour. An AEG system built using multi-task learning where the primary task is scoring the essay at the document-level, and learning the gaze behaviour is the auxiliary task at the word-level. Barrett *et al.* [2018] have shown this approach to benefit multiple NLP tasks like sentiment analysis, grammatical error detection and hate speech detection.

7 Conclusion

Gaze behaviour has been shown to aid multiple natural language processing tasks [Mishra and Bhattacharyya, 2018]. However, collecting gaze behaviour at run time is not feasible. Hence, in order to use gaze behaviour, we utilize different approaches, like multi-task learning, using type-aggregated values, etc.

In this paper, we first introduce the AI audience to different NLP tasks which are solved using gaze behaviour, like translation complexity, sarcasm understandability, text quality rating prediction, etc. We then discuss different tasks where we show that gaze behaviour aids in their solution. To solve any of these tasks, we require gaze behaviour data to be ready for training. In our paper, we also report gaze behaviour datasets created in multiple languages. Finally, we describe a pair of applications from the domain of education - complex word identification, and automatic essay grading - which could benefit a lot from using gaze behaviour based solutions.

References

[Alikaniotis *et al.*, 2016] Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. Automatic text scoring using neural networks. In *Proc. of the 54th ACL (Volume 1: Long Papers)*, pages 715–725. ACL, 2016.

[Attali and Burstein, 2006] Yigal Attali and Jill Burstein. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3), 2006.

[Barrett and Søgaard, 2015a] Maria Barrett and Anders Søgaard. Reading behavior predicts syntactic categories. In *Proceedings of the 19th CoNLL*, pages 345–349. Association for Computational Linguistics, 2015.

[Barrett and Søgaard, 2015b] Maria Barrett and Anders Søgaard. Using reading behavior to predict grammatical functions. In *Proc. of the 6th Workshop on Cognitive Aspects of Computational Language Learning*, pages 1–5. ACL, 2015.

[Barrett *et al.*, 2015] Maria Barrett, Željko Agić, and Anders Søgaard. The Dundee Treebank. In *The 14th International Workshop on Treebanks and Linguistic Theories (TLT 14)*, pages 242–248, 2015.

[Barrett *et al.*, 2016a] Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proc. of the 54th ACL (Volume 2: Short Papers)*, pages 579–584. ACL, 2016.

[Barrett *et al.*, 2016b] Maria Barrett, Frank Keller, and Anders Søgaard. Cross-lingual transfer of correlations between parts of speech and gaze features. In *Proc. of the 26th COLING*, pages 1330–1339. COLING, 2016.

[Barrett *et al.*, 2018] Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. Sequence classification with human attention. In *Proc. of the 22nd CoNLL*, pages 302–312. ACL, 2018.

[Berzak *et al.*, 2017] Yevgeni Berzak, Chie Nakamura, Suzanne Flynn, and Boris Katz. Predicting native language from gaze. In *Proc. of the 55th ACL (Volume 1: Long Papers)*, pages 541–551. ACL, 2017.

[Bingel *et al.*, 2018] Joachim Bingel, Maria Barrett, and Sigrid Klerke. Predicting misreadings from gaze in children with reading difficulties. In *Proc. of the 13th Workshop on NLP for Building Educational Applications (NLP-BEA)*, pages 24–34, 2018.

[Cheri *et al.*, 2016] Joe Cheri, Abhijit Mishra, and Pushpak Bhattacharyya. Leveraging annotators' gaze behaviour for coreference resolution. In *Proc. of the 7th Workshop on Cognitive Aspects of Computational Language Learning*, pages 22–26. ACL, 2016.

[Clarke and Lapata, 2006] James Clarke and Mirella Lapata. Models for sentence compression: A comparison across domains, training requirements and evaluation measures. In *Proc. of the 21st COLING and 44th ACL*, pages 377–384. ACL, 2006.

[Cop *et al.*, 2017] Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods*, 49(2):602–615, 2017.

[Dong and Zhang, 2016] Fei Dong and Yue Zhang. Automatic features for essay scoring – an empirical study. In *Proc. of the 2016 EMNLP*, pages 1072–1077. ACL, 2016.

[Dong *et al.*, 2017] Fei Dong, Yue Zhang, and Jie Yang. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proc. of the 21st CoNLL*, pages 153–162. ACL, 2017.

[Engbert *et al.*, 2005] Ralf Engbert, Antje Nuthmann, Eike M Richter, and Reinhold Kliegl. SWIFT: A dynamical model of saccade generation during reading. *Psychological review*, 112(4):777, 2005.

[Filippova *et al.*, 2015] Katja Filippova, Enrique Alfonseca, Carlos A. Colmenares, Lukasz Kaiser, and Oriol Vinyals. Sentence compression by deletion with LSTMs. In *Proc. of the 2015 EMNLP*, pages 360–368. ACL, 2015.

[Frank *et al.*, 2013] Stefan L Frank, Irene Fernandez Monsalve, Robin L Thompson, and Gabriella Vigliocco. Reading time data for evaluating broad-coverage models of english sentence processing. *Behavior Research Methods*, 45(4):1182–1190, 2013.

[González-Garduño and Søgaard, 2017] Ana Valeria González-Garduño and Anders Søgaard. Using gaze

to predict text readability. In *Proc. of the 12th Workshop on NLP for Building Educational Applications (NLP-BEA)*, pages 438–443. ACL, 2017.

[González-Garduno and Søgaard, 2018] Ana V González-Garduno and Anders Søgaard. Learning to predict readability using eye-movement data from natives and learners. In *Proc. of the 32nd AAAI*, 2018.

[Henderson and Ferreira, 1993] John M Henderson and Fernanda Ferreira. Eye movement control during reading: Fixation measures reflect foveal but not parafoveal processing difficulty. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 47(2):201, 1993.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[Hollenstein and Zhang, 2019] Nora Hollenstein and Ce Zhang. Entity recognition at first sight: Improving NER with eye movement information. In *Proc. of the 14th NAACL: HLT, Volume 1 (Long and Short Papers)*, pages 1–10. ACL, 2019.

[Hollenstein *et al.*, 2018] Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. ZuCo, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13, 2018.

[Joshi *et al.*, 2014] Aditya Joshi, Abhijit Mishra, Nivvedan Senthamilselvan, and Pushpak Bhattacharyya. Measuring sentiment annotation complexity of text. In *Proc. of the 52nd ACL (Volume 2: Short Papers)*, pages 36–41. ACL, 2014.

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

[Ke and Ng, 2019] Zixuan Ke and Vincent Ng. Automated essay scoring: A survey of the state of the art. In *Proc. of the 28th IJCAI*, pages 6300–6308. International Joint Conferences on Artificial Intelligence Organization, 2019.

[Kennedy *et al.*, 2003] Alan Kennedy, Robin Hill, and Joël Pynte. The Dundee Corpus. In *Proc. of the 12th European conference on eye movement*, 2003.

[Klerke and Plank, 2019] Sigrid Klerke and Barbara Plank. At a glance: The impact of gaze aggregation views on syntactic tagging. In *Proc. of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 51–61, 2019.

[Klerke *et al.*, 2015a] Sigrid Klerke, Sheila Castilho, Maria Barrett, and Anders Søgaard. Reading metrics for estimating task efficiency with MT output. In *Proc. of the 6th Workshop on Cognitive Aspects of Computational Language Learning*, pages 6–13, 2015.

[Klerke *et al.*, 2015b] Sigrid Klerke, Héctor Martínez Alonso, and Anders Søgaard. Looking hard: Eye tracking for detecting grammaticality of automatically compressed sentences. In *Proc. of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 97–105. Linköping University Electronic Press, Sweden, 2015.

[Klerke *et al.*, 2016] Sigrid Klerke, Yoav Goldberg, and Anders Søgaard. Improving sentence compression by learning to predict gaze. In *Proc. of the 12th NAACL: HLT*, pages 1528–1533. ACL, 2016.

[Kliegl *et al.*, 2004] Reinhold Kliegl, Ellen Grabner, Martin Rolfs, and Ralf Engbert. Length, frequency, and predictability effects of words on eye movements in reading. *European journal of cognitive psychology*, 16(1-2):262–284, 2004.

[Knight and Marcu, 2002] Kevin Knight and Daniel Marcu. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107, 2002.

[Kunze *et al.*, 2013] Kai Kunze, Hitoshi Kawaichi, Kazuyo Yoshimura, and Koichi Kise. Towards inferring language expertise using eye tracking. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 217–222. 2013.

[Landauer, 2003] Thomas K Landauer. Automated scoring and annotation of essays with the intelligent essay assessor. *Automated Essay Scoring: A Cross-disciplinary Perspective*, 2003.

[Laurinavichyute *et al.*, 2017] AK Laurinavichyute, Irina A Sekerina, SV Alexeeva, and KA Bagdasaryan. Russian sentence corpus: Benchmark measures of eye movements in reading in cyrillic. 2017.

[Li *et al.*, 2018] Xiangsheng Li, Yiqun Liu, Jiaxin Mao, Zexue He, Min Zhang, and Shaoping Ma. Understanding reading attention distribution during relevance judgement. In *Proc. of the 27th CIKM*, pages 733–742, 2018.

[Liang *et al.*, 2018] Guoxi Liang, Byung-Won On, Dongwon Jeong, Hyun-Chul Kim, and Gyu Sang Choi. Automated essay scoring: A siamese bidirectional LSTM neural network architecture. *Symmetry*, 10(12):682, 2018.

[Luke and Christianson, 2018] Steven G Luke and Kiel Christianson. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior research methods*, 50(2):826–833, 2018.

[Maas *et al.*, 2011] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proc. of the 49th ACL*, pages 142–150. ACL, 2011.

[Mak and Willems, 2019] Marloes Mak and Roel M Willems. Mental simulation during literary reading: Individual differences revealed with eye-tracking. *Language, Cognition and Neuroscience*, 34(4):511–535, 2019.

[Mathias *et al.*, 2018] Sandeep Mathias, Diptesh Kanojia, Kevin Patel, Samarth Agrawal, Abhijit Mishra, and Pushpak Bhattacharyya. Eyes are the windows to the soul: Predicting the rating of text quality using gaze behaviour. In *Proc. of the 56th ACL (Volume 1: Long Papers)*, pages 2352–2362. ACL, 2018.

[Matthies and Søgaard, 2013] Franz Matthies and Anders Søgaard. With blinkers on: Robust prediction of eye movements across readers. In *Proc. of the 2013 EMNLP*, pages 803–807. ACL, 2013.

[Mayfield and Rosé, 2013] Elijah Mayfield and Carolyn Penstein Rosé. LightSIDE: Open source machine learning for text. In *Handbook of Automated Essay Evaluation*, pages 146–157. Routledge, 2013.

[Mishra and Bhattacharyya, 2018] Abhijit Mishra and Pushpak Bhattacharyya. *Cognitively Inspired Natural Language Processing: An Investigation Based on Eye-tracking*. Springer, 2018.

[Mishra *et al.*, 2013] Abhijit Mishra, Pushpak Bhattacharyya, and Michael Carl. Automatically predicting sentence translation difficulty. In *Proc. of the 51st ACL (Volume 2: Short Papers)*, pages 346–351. ACL, 2013.

[Mishra *et al.*, 2016a] Abhijit Mishra, Diptesh Kanojia, and Pushpak Bhattacharyya. Predicting readers' sarcasm understandability by modeling gaze behavior. In *Proc. of the 30th AAAI*, pages 3747–3753, 2016.

[Mishra *et al.*, 2016b] Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. Harnessing cognitive features for sarcasm detection. In *Proc. of the 54th ACL (Volume 1: Long Papers)*, pages 1095–1104. ACL, 2016.

[Mishra *et al.*, 2016c] Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. Leveraging cognitive features for sentiment analysis. In *Proc. of the 20th CoNLL*. ACL, 2016.

[Mishra *et al.*, 2017] Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. Scanpath complexity: Modeling reading effort using gaze information. In *Proc. of the 31st AAAI*, pages 4429–4436, 2017.

[Mishra *et al.*, 2018] Abhijit Mishra, Srikanth Tamilselvam, Riddhiman Dasgupta, Seema Nagar, and Kuntal Dey. Cognition-cognizant sentiment analysis with multitask subjectivity summarization based on annotators' gaze behavior. In *Proc. of the 32nd AAAI*, 2018.

[Nicenboim *et al.*, 2016] Bruno Nicenboim, Pavel Logačev, Carolina Gattei, and Shravan Vasishth. When high-capacity readers slow down and low-capacity readers speed up: Working memory and locality effects. *Frontiers in Psychology*, 7:280, 2016.

[Nilsson and Nivre, 2009] Mattias Nilsson and Joakim Nivre. Learning where to look: Modeling eye movements in reading. In *Proc. of the 13th CoNLL*, pages 93–101. ACL, 2009.

[Paetzold and Specia, 2016] Gustavo Paetzold and Lucia Specia. SemEval 2016 task 11: Complex word identification. In *Proc. of the 10th SemEval*, pages 560–569. ACL, 2016.

[Page, 1966] Ellis B Page. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243, 1966.

[Pang and Lee, 2004] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proc. of the 42nd ACL*, pages 271–278, 2004.

[Parker *et al.*, 2017] Adam J Parker, Julie A Kirkby, and Timothy J Slattery. Predictability effects during reading in the absence of parafoveal preview. *Journal of Cognitive Psychology*, 29(8):902–911, 2017.

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

[Reichle *et al.*, 2003] Erik D Reichle, Keith Rayner, and Alexander Pollatsek. The EZ reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences*, 26(4):445–476, 2003.

[Safavi *et al.*, 2016] Molood S Safavi, Samar Husain, and Shravan Vasishth. Dependency resolution difficulty increases with distance in persian separable complex predicates: Evidence for expectation and memory-based accounts. *Frontiers in psychology*, 7:403, 2016.

[Søgaard, 2016] Anders Søgaard. Evaluating word embeddings with fMRI and eye-tracking. In *Proc. of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 116–121. ACL, 2016.

[Taghipour and Ng, 2016] Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In *Proc. of the 2016 EMNLP*, pages 1882–1891. ACL, 2016.

[Tay *et al.*, 2018] Yi Tay, Minh Phan, Luu Anh Tuan, and Siu Cheung Hui. SkipFlow: Incorporating neural coherence features for end-to-end automatic text scoring. In *Proc. of the 32nd AAAI*, pages 5948–5955. AAAI, 2018.

[Tjong Kim Sang and De Meulder, 2003] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proc. of the 7th CoNLL at NAACL-HLT 2003*, pages 142–147, 2003.

[Yaneva, 2016] Victoria Yaneva. Assessing text and web accessibility for people with autism spectrum disorder. 2016.

[Yimam *et al.*, 2018] Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. A report on the complex word identification shared task 2018. In *Proc. of the 13th Workshop on NLP for Building Educational Applications (NLP-BEA)*, pages 66–78. ACL, 2018.

[Zang *et al.*, 2018] Chuanli Zang, Ying Fu, Xuejun Bai, Guoli Yan, and Simon P Liversedge. Investigating word length effects in Chinese reading. *Journal of Experimental Psychology: Human Perception and Performance*, 44(12):1831–1841, 2018.

[Zhang and Litman, 2018] Haoran Zhang and Diane Litman. Co-attention based neural network for source-dependent essay scoring. In *Proc. of the 13th Workshop on NLP for Building Educational Applications (NLP-BEA)*, pages 399–409. ACL, 2018.