Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1712–1720
Marseille, 20-25 June 2022

The Copenhagen Corpus of Eye Tracking Recordings from Natural Reading of Danish Texts

Nora Hollenstein[1], Maria Barrett[2], Marina Björnsdóttir[1,2]
[1] Center for Language Technology, University of Copenhagen
[2] Computer Science Department, IT University of Copenhagen
nora.hollenstein@hum.ku.dk, mbarrett@itu.dk, marina.bjorns@gmail.com

Abstract

Eye movement recordings from reading are one of the richest signals of human language processing. Corpora of eye movements during reading of contextualized running text is a way of making such records available for natural language processing purposes. Such corpora already exist in some languages. We present CopCo, the Copenhagen Corpus of eye tracking recordings from natural reading of Danish texts. It is the first eye tracking corpus of its kind for the Danish language. CopCo includes 1,832 sentences with 34,897 tokens of Danish text extracted from a collection of speech manuscripts. This first release of the corpus contains eye tracking data from 22 participants. It will be extended continuously with more participants and texts from other genres. We assess the data quality of the recorded eye movements and find that the extracted features are in line with related research. The dataset available here: `https://osf.io/ud8s5/`.

Keywords: eye tracking, reading, Danish, corpus, psycholinguistics, cognitively-inspired NLP

1. Introduction

Records of eye movements during reading have been studied since decades in controlled psycholinguistic studies. The advantages of eye movement data for studying cognitive language processing are well established. For example, there is a thoroughly examined link between the fixation duration on a word and the cognitive effort required to process this word. During skilled reading, the eyes fixate sequentially through the text but approximately 10–15% of the fixations occur in a previously read part of the text to further process it. Most often, the reader is not aware of this. Therefore, eye movements allow us to study both early and later stages of cognitive text processing (Rayner et al., 1989).[1]

In controlled studies, participants typically read constructed individual sentences. Often, pairs of sentences with minimal differences are used to study one particular linguistic phenomenon (Brennan, 2016). However, as highlighted by Demberg and Keller (2019), there is a need for evaluating psycholinguistic theories on natural reading corpora rather than on highly infrequent sentence constructions specifically designed for the purpose of one experiment. Eye tracking and brain activity studies during unpaced reading of naturally occurring, contextualized text have recently gained more attention in the research community thanks to advances in recording technologies (Hamilton and Huth, 2018; Sato and Mizuhara, 2018). While natural eye tracking corpora cannot replace controlled psycholinguistic studies, it is a complementary method of studying some of the same phenomena with higher ecological validity. A naturalistic reading setup allows us to analyze real-time language processing from real-world text (Hasson and Egidi, 2015).

An obvious advantage of larger corpora of naturally occurring text is that the same corpus can be reused for multiple purposes. Researchers can either isolate the relevant linguistic phenomenon or model the entire reading process. Instead of comparing binarized groups of words, e.g., low-frequency and high-frequency words in a controlled study, eye tracking corpora allow us to model the entire spectrum as done by Kennedy et al. (2013). Similarly, controlled studies have shown that the syntax of a sentence is processed at the end of the sentence when comparing minimal pairs of sentences with different syntactical complexity (Traxler et al., 1997; Warren et al., 2009), the so-called wrap-up effect, but eye tracking corpora can answer how punctuation influences reading in contextualized naturally occurring sentences that contain the full spectrum of simple and more complex syntactic constructions (Pynte and Kennedy, 2007).

Another benefit of the larger natural eye movement corpora is their potential for NLP (Hollenstein et al., 2020a). Some of these corpora are large enough to train cognitively inspired models. The growing list of gaze-augmented NLP models includes tasks such as image captioning (Takmaz et al., 2020), named entity recognition (Hollenstein and Zhang, 2019), sentiment analysis (Mishra et al., 2016b), or part-of-speech tagging (Barrett et al., 2016).[2] More recently, eye tracking data from reading has also been leveraged to evaluate and interpret computational language models (Sood et al., 2020; Abdou et al., 2019; Hollenstein and Beinborn,

[1] For a review, see Clifton et al. (2007).

[2] See Mathias et al. (2020) and Barrett and Hollenstein (2020) for more extensive reviews.

Speech ID	Sents.	Tokens	Types	Sent. length	Token length	Freq.	LIX
1125	132	1917	611	14.52 (8.75, 1-45)	3.84 (2.87, 1-22)	0.74	30.16
1165	71	1319	454	18.58 (11.5, 1-59)	3.71 (2.39, 1-17)	0.78	30.12
1317	107	1830	683	17.1 (7.53, 5-44)	4.73 (3.86, 1-34)	0.74	43.19
1318	114	2143	711	18.8 (9.73, 5-51)	4.38 (3.39, 1-25)	0.75	41.49
1323	100	2044	720	20.44 (10.59, 4-50)	4.62 (3.7, 1-25)	0.74	44.52
7797	119	1639	645	13.77 (8.63, 1-37)	3.94 (2.96, 1-24)	0.70	31.00
7856	59	2139	634	36.25, (26.47, 2-139)	3.75 (2.55, 1-21)	0.77	45.58
7905	106	2648	1056	24.98 (14.74, 3-67)	4.5, (3.3, 1-24)	0.73	48.58
7946	134	1700	591	12.69 (9.72, 1-80)	3.82 (2.74, 1-20)	0.71	26.43
10365	126	1782	615	14.14, (8.05, 2-43)	3.69 (2.64, 1-24)	0.71	26.77
10440	89	1150	409	12.92, (7.64, 2-38)	3.98 (3.23, 1-22)	0.73	29.08
11171	51	1170	477	22.94 (13.78, 1-56)	3.97 (2.86, 1-20)	0.73	39.09
12063	57	1172	498	20.56 (14.26, 2-57)	3.89 (2.81, 1-23)	0.68	34.79
17526	109	2846	784	26.11 (15.91, 1-71)	3.8, (2.85, 1-24)	0.74	39.79
18473	37	978	391	26.43 (13.98, 4-58)	4.51 (3.56, 1-20)	0.75	50.14
18561	82	1260	410	15.37 (7.91, 1-37)	3.87, (2.93, 1-18)	0.74	31.70
18670	81	1282	480	15.83 (9.38, 1-54)	4.05 (3.09, 1-21)	0.72	34.81
22811	54	1357	524	25.13 (19.4, 1-97)	4.16 (3.41, 1-29)	0.73	41.49
26670	102	2215	641	21.72 (11.7, 1-52)	3.94 (2.73, 1-22)	0.79	37.79
26682	119	2306	737	19.38 (10.66, 1-52)	4.1, (2.93, 1-21)	0.78	37.85
total	1832	34897	5872	19.05 (13.07, 1-139)	4.07 (3.08, 1.34)	0.74	37.22

Table 1: Dataset statistics. The speech ID is the original ID from the source corpus; sentence length is the mean number of tokens per sentence (with standard deviation and range in brackets); token length is the mean number of characters per token (with standard deviation and range in brackets); frequency is the proportion of words included in the 10,000 most common Danish words (Source: `https://korpus.dsl.dk/resources/details/freq-lemmas.html`); LIX is the readability score as described in Section 3.1.

2021).

There are eye tracking corpora from natural reading available in other languages, but as of yet, no such resource is available for Danish. We present CopCo, the Copenhagen Eye Tracking Corpus. It is the first Danish eye tracking corpus with contextualized, running text and self-paced reading. CopCo is a growing dataset, and the first release includes recordings from 22 participants over more than 30,000 tokens. CopCo is freely available here: `https://osf.io/ud8s5/`.

2. Related Work

Some of the existing corpora emerged from psycholinguistic experiments, while others were tailored for their use in natural language processing applications. In English, there are a number of eye tracking corpora of skilled adult readers performing self-paced reading of contextualized text, e.g., Kennedy et al. (2003), Cop et al. (2017), Luke and Christianson (2018), Mishra et al. (2016a). Some encompass more than 50,000 tokens, while others are smaller and focus on individual sentence processing (Frank et al., 2013; Hollenstein et al., 2018; Hollenstein et al., 2020b). Several participants read the same text; in the Provo Corpus as many as 470 (Luke and Christianson, 2018), but in the remaining cited corpora, the number of subjects reading the same text is between 10 and 20.

Natural eye tracking corpora also exist in other languages. Some of these are again smaller and focus on individual sentence processing: Husain et al. (2015) for Hindi, Safavi et al. (2016) for Persian, Laurinavichyute et al. (2019) for Russian, and Pan et al. (2021) for Chinese. The most recent release is the Multilingual Eye Movement Corpus (MECO; (Kuperman et al., 2020; Siegelman et al., 2022)), a resource that includes parallel data from 580 readers in 13 different languages, reading in their native language as well as in English, following the same experiment protocol. However, this corpus does not include Danish. Others study specific linguistic aspects, for example, Cop et al. (2017) study bilingual reading processing of Dutch and English on a full novel, Jäger et al. (2021) analyze reading patterns between experts and non-experts.

Finally, some of the existing resources are explicitly targeted for their use in NLP applications. For instance, Yi et al. (2020) compile a Chinese dataset of gaze behavior from text summarization and Sood et al. (2021) provide an English dataset of visual question answering.

CopCo is a new resource in this landscape of eye tracking corpora and provides data to analyze psycholinguistic research questions as well as NLP applications. All materials are freely available so that annotations or labels for specific NLP tasks can be added in future work.

ID	Age	Sex	Comp. score	# Speeches	# Questions	Reading time
P01	29	F	0.92	4	13	25.95
P02	62	F	0.83	6	18	14.28
P03	23	F	0.95	6	20	18.66
P04	26	F	0.88	2	8	20.37
P05	44	F	0.85	7	26	13.88
P06	47	M	0.78	6	18	14.68
P07	26	F	0.81	4	16	22.27
P08	32	M	0.8	2	5	14.72
P09	39	F	0.0	1	1	21.25
P10	25	F	0.81	4	16	22.41
P11	32	F	0.78	6	23	13.62
P12	29	M	0.86	4	14	18.27
P13	32	M	1.0	3	8	15.36
P14	59	F	0.85	6	20	13.19
P15	25	F	0.87	4	15	19.13
P16	22	M	0.79	4	14	14.24
P17	37	F	1.0	2	7	17.75
P18	26	F	0.82	4	11	18.50
P19	21	F	0.62	4	13	16.19
P20	24	F	0.87	6	23	10.03
P21	26	F	1.0	4	16	14.41
P22	23	F	0.80	4	15	18.67
mean	32	-	0.82	4.13	14.3	17.27

Table 2: Participant statistics. Comp. score: proportion of correctly answered reading comprehension questions; number of speeches and number of questions read/answered by this participant; absolute reading time: seconds spent on each screen.

3. Experiment Design

3.1. Reading Materials

All reading materials are Danish speech manuscripts from `https://dansketaler.dk/`. We selected speeches limited to the following categories: event or conference speeches, speeches given by a bonfire at a solstice event, and high school graduation speeches. None of the conferences are scientific conferences for a highly specialised audience. The speeches from the selected categories are expected to be engaging for a general audience. Although speech manuscripts are not a frequently studied genre, the Danske Taler archive provides a great resource of longer, yet self-contained Danish texts of current language use without copyright. The content is largely comparable to essays; a piece of writing on a particular topic, often from a personal point of view. We sampled speeches from the years 2010–2019. This selection returned 46 speeches, and we manually subsampled them to remove speeches that are not interesting to a general audience, as well as to get a broad distribution of speaker demographics in the final sample. We balanced the gender distribution of the speakers such that the corpus contains ten male and ten female speakers.

Furthermore, our objective was to obtain as broad an age range and range of the geographical location of the speech event as possible. Some of the speeches were already proofread by Danske Taler. Nevertheless, a native Danish speaker proofread all speeches chosen for this data collection.[3]

In total, CopCo contains 34,897 tokens in 1,832 sentences in a selection of 20 speeches. Table 1 presents the statistics of the data set for each speech, including the number of words, sentences, the average word length, sentence length, and the LIX score (Björnsson, 1968) as calculated by the `readcalc` Python library.[4] The LIX score is a simple readability metric considering the length of words and the length of sentences. A score of 25–34 is considered an easy text for skilled adult readers, and >55 is considered difficult.

3.2. Comprehension Questions

After approximately 20% of all paragraphs longer than 100 characters, the subjects are presented with a comprehension question related to the previously read paragraph to prevent mindless reading. The subset of para-

[3]Not all speech manuscripts used canonical punctuation but rather marked pauses for the speaker - sometimes with hyphens. We tried to maintain the texts in their original form as far as possible. We edited clearly incorrect punctuation but did allow extensive use of hyphen instead of comma and full stop. We acknowledge that replacing full stops with hyphens inaccurately inflates the readability score.

[4]`https://pypi.org/project/ReadabilityCalculator/`

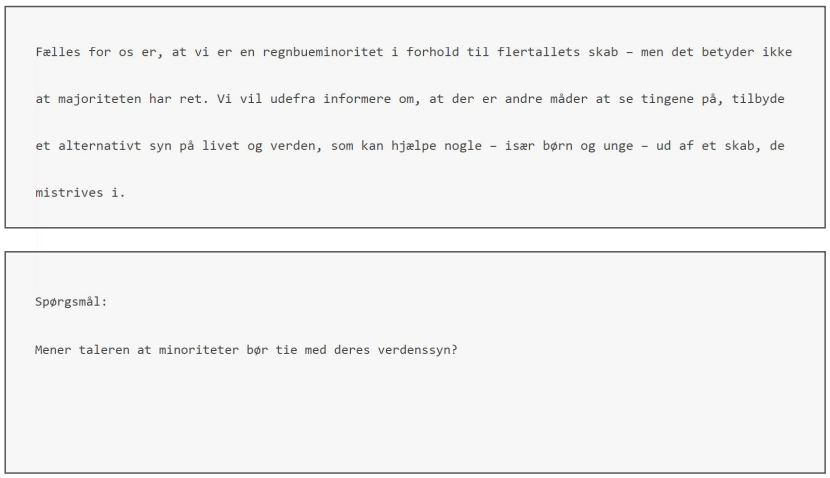

Figure 1: Example screen of a text paragraph and the following question on the next screen.

graphs was selected randomly. Each question must be answered with *yes* or *no*. There is no option to skip a question. The questions either ask about the factual details described in the text or a verification of a more general and shallow synthesis of the text. In total, there are 68 questions – 30 where the correct answer is *no*, 38 where the correct answer is *yes*.

3.3. Participants

The participants are adult, native speakers of Danish. All have normal vision or corrected-to-normal vision (glasses or contact lenses) and no known reading impairments. At the time of publication, we collected eye movement data from 22 participants (23% male, between 21 and 62 years old, with the highest completed education levels ranging between high school and PhD). Participant recruitment is still in progress and data of further participants will be released as soon as available. Participation was rewarded with a symbolic gift. All participants gave written consent to their participation and the reuse of the data for research purposes prior to the start of the experiment. This study was approved by the Research Ethics Committee at the Faculty of Humanities of the University of Copenhagen.

Table 2 shows the details of the participant population. For each participant, we report age and sex, the number of speeches they read, the number of comprehension questions answered, the average reading time per screen, as well as the proportion of correctly answered comprehension questions. In total, the participants have read 95 texts. Each text has been read by 3 to 8 readers. Due to bad calibration and technical problems, the data of participant P14 are not used for any further analysis.

3.4. Recording Procedure

At the beginning of the experiment, participants were instructed to move as little as possible and read as naturally as possible, as they would read for comprehension outside the laboratory. Participants rested their heads on an adjustable chin rest to limit head and body movements, and they used a control pad to move to the next screen and answer the comprehension questions in their own speed. Reverting to a previous screen was not possible. The reading was self-paced, which means that participants pressed a key after finishing reading each screen to move onto the next. There was no time restriction; neither to the reading of a single screen nor to the session duration. Instructions for the task were presented orally as well as on the computer screen before the experiment start. All participants were first presented with a short speech as a practice round.

The experiment was split into blocks of two speeches. The order of the blocks and the order of the speeches within a block were randomized. The experiment design allows a flexible length of the recording sessions. Each session entails at least one block (i.e., two speeches) and if the participant is not too tired, subsequent blocks can be added. On average, participants read 4.13 speeches per session. This setup allows us to extend the corpus with additional texts in the future while maintaining consistency in the experiment procedure.

3.5. Stimulus Presentation

The text passages presented on each screen resembled the author's original division of the story into paragraphs as much as possible. Comprehension questions were presented on separate screens and clearly marked with the title "Spørgsmål" (translation: "Question").

The text was presented in a black, monospaced font (font type: Consolas; font size: 16) on a light-gray background (RGB: 248,248,248) as shown in Figure 1. The texts spanned multiple lines (max. 10) with triple line spacing. The text was presented with a 140 pixels margin at the top and bottom, and 200 pixels on the left and right.

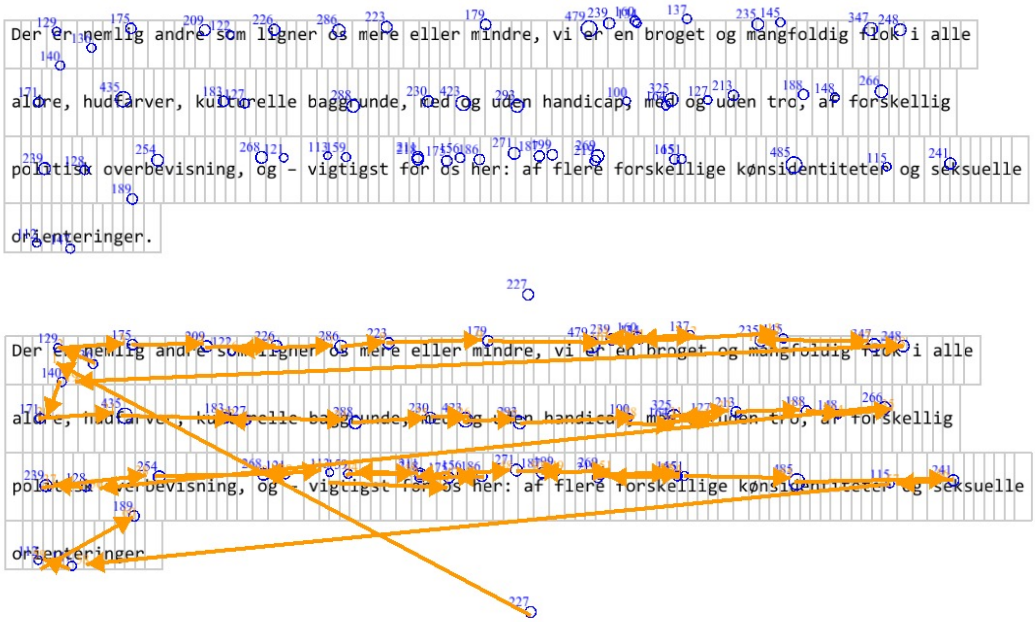

Figure 2: (Top) Fixations detected from a single reader; the numbers in blue show the fixation duration in milliseconds. (Bottom) The corresponding saccades between the fixations, the arrow heads mark the direction of the eye movement.

4. Data Acquisition

Eye movement data was collected with an infrared video-based EyeLink 1000 Plus eye tracker (SR Research). The experiment was designed with the SR Experiment Builder software. Data is recorded with a sampling rate of 1000 Hz.

Participants were seated at a distance of approximately 85 cm from a 27-inch monitor (display dimensions 590 x 335 mm, resolution 1920 x 1080 pixels). We recorded monocular eye tracking data of the right eye. In a few cases of calibration difficulties, the left eye was tracked.

A 9-point calibration was performed at the beginning of the experiment. The calibration was validated after each block. Re-calibration was conducted if the quality was not good (worst point error $< 1.5°$, average error $< 1.0°$). Drift correction was performed after each text passage.

5. Preprocessing

In this Section, we describe the processing from the raw data recordings to the extraction of character-level and word-level reading time features. We share the following versions of the data: raw, character-level fixation information, character-level saccade information, and word-level eye tracking features.

5.1. Gaze Event Detection

A fixation is defined as a time window during which the eye is relatively still and focuses on the same point. For CopCo, the eye movement events are generated in real-time by the EyeLink eye tracker software during recording with a velocity- and acceleration-based saccade detection method. A fixation event is defined by the algorithm as any period that is not a saccade or a blink. Hence, the raw data consist of (x,y) gaze location coordinates for individual fixations.

We use the DataViewer software by SR Research to extract fixation events for all areas of interest. Areas of interest are automatically defined as rectangular boxes that surround each individual character of a text on the screen, as shown in Figure 2. For later analysis, only fixations within the boundaries of each displayed character are extracted. Therefore, data points distinctly not associated with reading are excluded. An example of the resulting fixations and saccades is shown in Figure 2.

5.2. Feature Extraction

In a second step, we use custom Python code to map and aggregate character-level features to word-level features. Figure 3 depicts this process. To the best of our knowledge, we are the first to share such an automatic conversion script.[5]

We extract the following eye tracking features:

1. *Number of fixations*, the total amount of fixations on the current word, including all passes.

2. *First fixation duration*, the duration (in milliseconds) of the first fixation on the prevailing word.

[5]The code is available here: `https://github.com/norahollenstein/copco-processing`

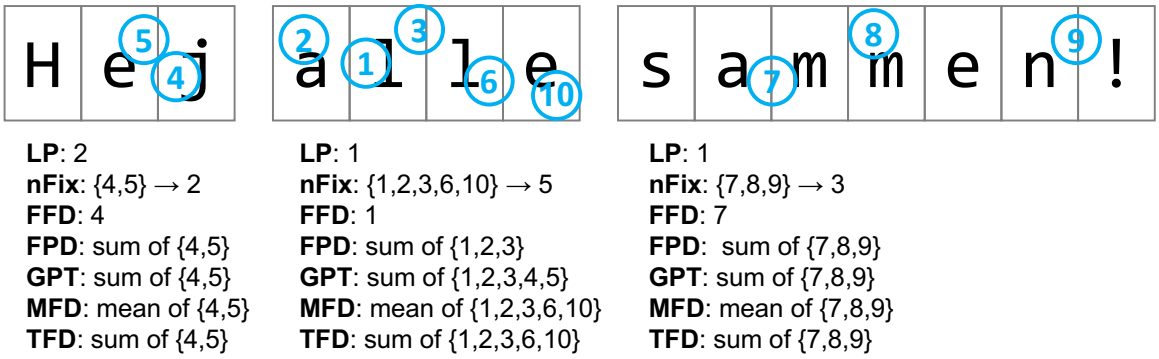

Figure 3: Character-level to word-level feature mapping. The blue circles represent individual fixations and their order, and the gray boxes mark the character-level interest areas. Abbreviations: Landing position (LP), number of fixations (nFix), first fixation duration (FFD), first pass duration (FPD), go-past time (GPT), total fixation duration (TFD), mean fixtion duration (MFD).

3. *Mean fixation duration*, the sum of all fixation durations (in milliseconds) on the current word divided by the number of fixations.

4. *Total fixation duration*, the sum of all fixation durations (in milliseconds) on the current word.

5. *First pass duration*, the summed duration (in milliseconds) of all fixations on the current word prior to progressing out of the current word (to the left or right).

6. *Go-past time*, the sum duration (in milliseconds) of all fixations prior to progressing to the right of the current word, including regressions to previous words that originated from the current word.

7. *Landing position*, index of the first character fixation in the prevailing word (if the fixation falls into multiple interest areas, the character on the left is chosen).

8. *Mean saccade duration*, the mean duration (in milliseconds) of all saccades originating from the current word.

9. *Peak saccade velocity*: Maximum gaze velocity (in visual degrees per second) of all saccades originating from the current word.

These features are defined to cover the reading process from early lexical access to later syntactic integration. The selection of features is inspired by similar corpora in other languages (Hollenstein et al., 2018; Cop et al., 2017) and extended to include character-level features, which will enable more fine-grained psycholinguistic analyses (i.e., differences in landing positions and character types between languages and scripts).

6. Data Validation

To ensure the quality of the recorded data, we present a series of analyzes that take a closer look at reading comprehension, the effects of word length and word frequency, the effects on landing position at character level, and a comparison of the extracted features.

Reading comprehension. Based on the scores of the reading comprehension questions of all participants (as presented in Table 2), the mean accuracy is 82%, with a minimum of 62% and a maximum of 100%. Therefore, no participant data needs to be excluded due to low comprehension. On average, participants read four speeches in a one-hour session. The mean reading time per screen is 17.65 seconds (± 3.84). All except one participants fall within two standard deviations of the mean. In this participant population, there is no significant correlation between the reading comprehension scores and the reading times per screen (Spearman's rank correlation coefficient: 0.1, $p > 0.6$).

Word length & word frequency. Eye movements during reading are regulated by various lexical aspects such as word length and word frequency: Longer and less frequent words are more likely to be fixated. Furthermore, these word characteristics affect fixation duration similarly across languages, but the size of the effect depends on the language and the script (Laurinavichyute et al., 2019; Bai et al., 2008). As we can observe in the new CopCo corpus, this is also the case in Danish. In Figure 4, we show the effect of word length found in the eye tracking data recorded from the Danish stimulus. Figure 5 shows how more frequent words in the Danish language are skipped more often in the recorded eye tracking data. The general skipping rate for participants (i.e., the proportion of words that are not fixated) lies between 0.29 and 0.58.

Landing position. Next, we analyze the landing position within a word. In early research, Liversedge and Underwood (1998) suggested that orthographic information, such as the frequency of characters visible in the parafovea, may influence the landing position on the following word. In the CopCo data, we find that the number of times a character is fixated first within

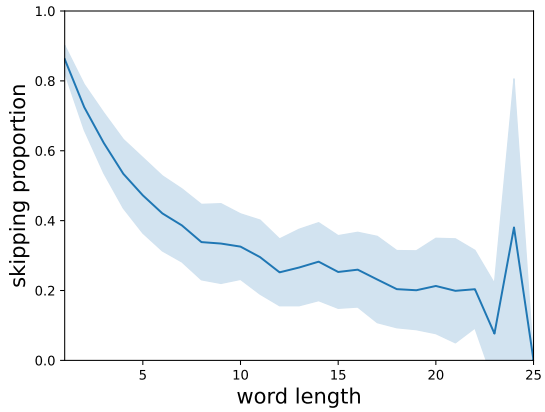

Figure 4: Effect of word length on the skipping proportion across all participants (i.e., the proportion of readers that fixate a given word), with the standard deviation in the shaded area.

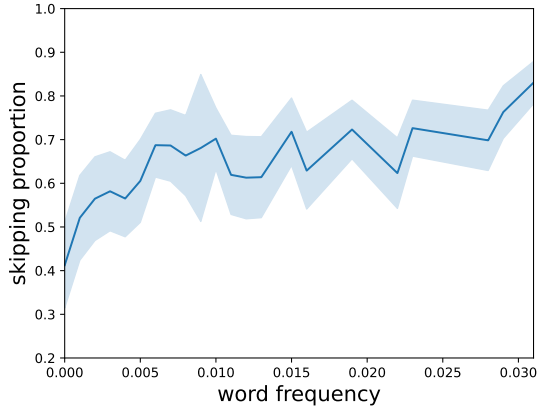

Figure 5: Effect of word frequency on the skipping proportion across all participants (i.e., the proportion of readers that fixate a given word), with the standard deviation in the shaded area.

a word correlates only moderately with the character frequency in Danish (Spearman's rank correlation coefficient: 0.35, $p = 0.069$). Rayner and Morris (1992) concluded that low-level visual information (primarily word length) is the key determinant of the initial landing position on a word during reading. We find this effect in the CopCo data: The character landing position index highly correlates with the word length, meaning that for longer words the gaze tends to land on later characters (Spearman's rank correlation coefficient: 0.90, $p < 0.0001$). Finally, the fixation duration on the landing character does not differ significantly between vowels and consonants.

Feature ranges. Lastly, we compare the extracted word-level eye tracking features to existing corpora. Fixations shorter than 100 ms were excluded from the analysis, because these are unlikely to reflect fixations

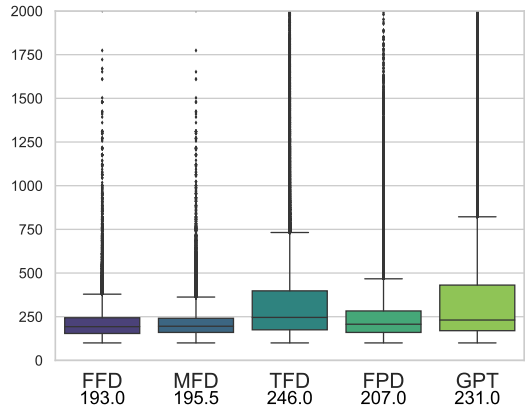

Figure 6: Feature ranges for the first fixation duration (FFD), mean fixation duration (MFD), total fixation duration (TFD), first-pass duration (FPD), and go-past time (GPT) with the median feature value below each boxplot.

relevant for reading (Sereno and Rayner, 2003). On average, a word is fixated 0.69 times (standard deviation: 1.05). Figure 6 shows the mean and range of each of the five most commonly extracted fixation features: first fixation duration, mean fixation duration, total fixation duration, first-pass duration and go-past time. These values are in line with similar corpora in other languages, e.g., the GECO corpus (Cop et al., 2017) and the ZuCo corpus (Hollenstein et al., 2018).

7. Outlook

CopCo can be used for applications in the fields of natural language processing as well as in psycholinguistics and reading research. In NLP, eye movement features can be leveraged to improve models for syntactic and semantic language understanding tasks (e.g., part-of-speech tagging (Barrett et al., 2016; Klerke and Plank, 2019) or named entity recognition (Hollenstein and Zhang, 2019)). This new eye tracking dataset also allows us to analyze and interpret language models or task-specific NLP models. For example, we can investigate machine-learning based explainability mechanisms such as attention and saliency in Danish language models, as suggested by Hollenstein and Beinborn (2021) or Sood et al. (2020) for English.

In psycholinguistics, the CopCo data can be used to study human reading, including the analysis of reading patterns of Danish native speakers, investigating differences between individual readers or subgroups of readers (e.g., split by age or gender), and the prediction of eye movements from reading Danish texts to develop more accurate reading models. Moreover, it enables further exploration of linguistic phenomena in natural reading, e.g., processing of relative clauses or negation. Finally, this new Danish eye tracking corpus allows cross-linguistic analysis of eye movements

while reading by comparing with available eye tracking corpora in other languages.

CopCo is designed to be a growing corpus. Future releases will include (i) recordings from additional participant populations such as dyslexic readers and Danish language learners, and (ii) reading materials of other text genres, for instance, Wikipedia articles or social media posts.

Acknowledgements

Maria Barrett is supported by a research grant (34437) from VILLUM FONDEN.

8. Bibliographical References

Abdou, M., Kulmizev, A., Hill, F., Low, D. M., and Søgaard, A. (2019). Higher-order comparisons of sentence encoder representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5842–5848.

Bai, X., Yan, G., Liversedge, S. P., Zang, C., and Rayner, K. (2008). Reading spaced and unspaced Chinese text: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, 34(5):1277.

Barrett, M. and Hollenstein, N. (2020). Sequence labelling and sequence classification with gaze: Novel uses of eye-tracking data for natural language processing. *Language and Linguistics Compass*, 14(11):1–16.

Barrett, M., Bingel, J., Keller, F., and Søgaard, A. (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.

Björnsson, C. (1968). Läsbarhet, liber. *Stockholm, Sweden*.

Brennan, J. (2016). Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass*, 10(7):299–313.

Clifton, C., Staub, A., and Rayner, K. (2007). Eye movements in reading words and sentences. In *Eye Movements*, pages 341–371. Elsevier.

Cop, U., Dirix, N., Drieghe, D., and Duyck, W. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49(2):602–615.

Demberg, V. and Keller, F. (2019). Cognitive models of syntax and sentence processing. *Human Language: From Genes and Brains to Behavior; Hagoort, P., Ed*, pages 293–312.

Frank, S. L., Monsalve, I. F., Thompson, R. L., and Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45(4):1182–1190.

Hamilton, L. S. and Huth, A. G. (2018). The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, pages 1–10.

Hasson, U. and Egidi, G. (2015). What are naturalistic comprehension paradigms teaching us about language? In *Cognitive Neuroscience of Natural Language Use*, pages 228–255. Cambridge University Press.

Hollenstein, N. and Beinborn, L. (2021). Relative importance in sentence processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 141–150.

Hollenstein, N. and Zhang, C. (2019). Entity recognition at first sight: Improving NER with eye movement information. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1–10.

Hollenstein, N., Rotsztejn, J., Troendle, M., Pedroni, A., Zhang, C., and Langer, N. (2018). ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*.

Hollenstein, N., Barrett, M., and Beinborn, L. (2020a). Towards best practices for leveraging human language processing signals for natural language processing. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pages 15–27.

Hollenstein, N., Troendle, M., Zhang, C., and Langer, N. (2020b). ZuCo 2.0: A Dataset of Physiological Recordings During Natural Reading and Annotation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 138–146.

Husain, S., Vasishth, S., and Srinivasan, N. (2015). Integration and prediction difficulty in Hindi sentence comprehension: Evidence from an eye-tracking corpus. *Journal of Eye Movement Research*, 8(2).

Jäger, L., Kern, T., and Haller, P. (2021). Potsdam Textbook Corpus (PoTeC): Eye tracking data from experts and non-experts reading scientific texts. available on OSF, DOI 10.17605/OSF.IO/DN5HP.

Kennedy, A., Hill, R., and Pynte, J. (2003). The Dundee corpus. In *Proceedings of the 12th European Conference on Eye Movement*.

Kennedy, A., Pynte, J., Murray, W. S., and Paul, S.-A. (2013). Frequency and predictability effects in the Dundee Corpus: An eye movement analysis. *The Quarterly Journal of Experimental Psychology*, 66(3):601–618.

Klerke, S. and Plank, B. (2019). At a glance: The impact of gaze aggregation views on syntactic tagging. In *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 51–61, Hong Kong, China, November. Association for Computational Linguistics.

Kuperman, V., Siegelman, N., Schroeder, S., Alexeeva, A., Acartürk, C., Amenta, S., Bertram, S., Bonandrini, R., Brysbaert, M., Chernova, D., et al. (2020). Text reading in English as a second language: Evidence from the multilingual eye-movements corpus (MECO). *Studies in Second Language Acquisition*.

Laurinavichyute, A., Sekerina, I. A., Alexeeva, S., and Bagdasaryan, K. (2019). Russian Sentence Corpus: Benchmark measures of eye movements in reading in Cyrillic. *Behavior Research Methods*, 51(3):1161–1178.

Liversedge, S. P. and Underwood, G. (1998). Foveal processing load and landing position effects in reading. In *Eye Guidance in Reading and Scene Perception*, pages 201–221. Elsevier.

Luke, S. G. and Christianson, K. (2018). The Provo corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, pages 1–8.

Mathias, S., Kanojia, D., Mishra, A., and Bhattacharyya, P. (2020). A survey on using gaze behaviour for natural language processing. *Proceedings of IJCAI*.

Mishra, A., Kanojia, D., and Bhattacharyya, P. (2016a). Predicting readers' sarcasm understandability by modeling gaze behavior. In *AAAI Conference on Artificial Intelligence*, pages 3747–3753.

Mishra, A., Kanojia, D., Nagar, S., Dey, K., and Bhattacharyya, P. (2016b). Leveraging cognitive features for sentiment analysis. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 156–166.

Pan, J., Yan, M., Richter, E. M., Shu, H., and Kliegl, R. (2021). The Beijing Sentence Corpus: A Chinese sentence corpus with eye movement data and predictability norms. *Behavior Research Methods*, pages 1–12.

Pynte, J. and Kennedy, A. (2007). The influence of punctuation and word class on distributed processing in normal reading. *Vision Research*, 47(9):1215–1227.

Rayner, K. and Morris, R. K. (1992). Eye movement control in reading: Evidence against semantic preprocessing. *Journal of Experimental Psychology: Human perception and performance*, 18(1):163.

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

Safavi, M. S., Husain, S., and Vasishth, S. (2016). Dependency resolution difficulty increases with distance in persian separable complex predicates: Evidence for expectation and memory-based accounts. *Frontiers in Psychology*, 7:403.

Sato, N. and Mizuhara, H. (2018). Successful encoding during natural reading is associated with fixation-related potentials and large-scale network deactivation. *Eneuro*, 5(5).

Sereno, S. C. and Rayner, K. (2003). Measuring word recognition in reading: Eye movements and event-related potentials. *Trends in Cognitive Sciences*, 7(11):489–493.

Siegelman, N., Schroeder, S., Acartürk, C., Ahn, H.-D., Alexeeva, S., Amenta, S., Bertram, R., Bonandrini, R., Brysbaert, M., Chernova, D., et al. (2022). Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). *Behavior Research Methods*, pages 1–21.

Sood, E., Tannert, S., Frassinelli, D., Bulling, A., and Vu, N. T. (2020). Interpreting attention models with human visual attention in machine reading comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 12–25.

Sood, E., Kögel, F., Strohm, F., Dhar, P., and Bulling, A. (2021). VQA-MHUG: A gaze dataset to study multimodal neural attention in visual question answering. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 27–43.

Takmaz, E., Pezzelle, S., Beinborn, L., and Fernández, R. (2020). Generating image descriptions via sequential cross-modal alignment guided by human gaze. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4664–4677.

Traxler, M. J., Bybee, M. D., and Pickering, M. J. (1997). Influence of connectives on language comprehension: eye tracking evidence for incremental interpretation. *The Quarterly Journal of Experimental Psychology Section A*, 50(3):481–497.

Warren, T., White, S. J., and Reichle, E. D. (2009). Investigating the causes of wrap-up effects: Evidence from eye movements and E–Z Reader. *Cognition*, 111(1):132–137.

Yi, K., Guo, Y., Jiang, W., Wang, Z., and Sun, L. (2020). A dataset for exploring gaze behaviors in text summarization. In *Proceedings of the 11th ACM Multimedia Systems Conference*, pages 243–248.