The Ghent Eye Tracking COrpus

Nicolas Dirix[a], and Wouter Duyck[a,b], [a] Ghent University, Ghent, Belgium and [b] The Accreditation Organization of the Netherlands and Flanders (NVAO), Den Haag, the Netherlands

Abstract

The Ghent Eye tracking Corpus (GECO) is a collection of eye tracking data from participants who read an entire novel, presented in paragraphs on a screen. It includes data from Dutch–English bilinguals, who read one-half of the novel in their first language and the other in their second language, and English monolinguals. The corpus contains approximately 1.75 million datapoints and has been utilized in the field of psycholinguistics in diverse topics such as (bilingual) visual word recognition and sentence processing. GECO is thus a valuable resource for advancing our understanding of the underlying cognitive processes of reading behavior.

Key Points

- GECO is a corpus of natural reading of an entire novel, facilitating eye movement research on monolingual and bilingual reading.
- Eye tracking provides a wide range of measures for analyzing reading behavior, including spatial and temporal variables, offering insights into underlying cognitive processes.
- Corpora like GECO allow for the study of natural reading processes, offering diverse stimuli and enhancing statistical power and enabling investigations at various levels of language processing.
- The wide applicability of GECO is supported by its use in a wide range of studies in visual word recognition, sentence processing, statistical learning, and modeling of reading processes.
- GECO expansions include GECO-CN for Chinese-English bilinguals and GECO-MT comparing reading of human and machine translated text.

Introduction

GECO (**Cop et al., 2017**) is an eye movement corpus of monolingual and bilingual reading. In the following article, we briefly highlight the antecedents of GECO and provide insights in the methodology and resulting data. Furthermore, as proof of concept of its potential applicability, we present a range of studies from various domains in psycholinguistics that applied GECO data for empirical research.

Eye Tracking

Eye tracking is a technique in which eye movements are precisely captured. More specifically, the location and duration of fixations (when the eyes rest on a rest on a certain position) and saccades (when the eye jumps between fixations) are registered. In the past 50 years, eye tracking made increasingly important contributions to empirical research and theoretical advancement in psycholinguistics (**Clifton et al., 2016**; **Rayner, 1998**).

Eye tracking is widely used due to its substantial advantages. First, state-of-the-art eye trackers offer exceptional spatial and temporal resolution, capturing the position of a fixation up to 2000 times per second, at word or even letter level. Second, an impressive range of variables can be constructed from eye movement data, such as the first fixation duration or the total reading time of a word, the order of fixations, whether regressions were made, … A third advantage is that this wide range of measures allows for a fine-grained analysis of the reading process. Researchers assume that eye movements are related to underlying cognitive processes (i.e., the so-called eye-mind hypothesis; see **Brysbaert & Drieghe, 2024**, for a recent discussion). Some measures are linked to earlier processes such as lexical access (e.g., the first fixation duration), others to later processes such as the integration of a word in the sentence (e.g., number of regressions; **Boston et al., 2008**). Finally, eye-tracking can be applied in more controlled experimental settings, as well as in so-called “natural reading”: reading of sentences or text without any further instructions or tasks. Natural reading typically has a higher degree of ecological validity than controlled experiments, as it mimics real-life reading.

Drawbacks of eye tracking include the need of expensive equipment, its time-consuming nature (typically one participant at a time), and several practical issues (e.g., exclusion of participants with lower visual acuity, erroneous recordings because of head movements).

Eye Tracking Corpora

Extensive linguistic corpora have a critical role in expanding our understanding of the cognitive processes underlying language comprehension. While conventional experimental setups focus on specific hypotheses with carefully chosen stimuli, larger and more diverse datasets are required to evaluate the real-world applicability of computational models of reading (**Dijkstra et al., 2019**; **Snell, van Leipsig, et al., 2018**) or eye movement control (e.g., **Reichle et al., 2003**).

Corpus studies include expansive samples of unselected stimuli, offering heightened statistical power and greater representativeness. Notably, large-scale lexical decision projects like the English Lexicon Project (ELP; **Balota et al., 2007**) have provided invaluable insights into large-scale empirical validation and further exploration of various psycholinguistic phenomena (e.g., **Brysbaert et al., 2018**; **Kuperman & Van Dyke, 2013**), as well as further advancement of computational models (e.g., **Chang et al., 2019**).

Similarly, eye tracking corpora of natural reading enable researchers to address contextualized reading processes by means of fine-grained analyses of the time course of the reading process (**Brysbaert & Drieghe, 2024**; **Cop et al., 2017**). For instance, the Dundee Corpus (**Kennedy & Pynte, 2005**) consists of eye movement data of participants reading newspaper articles. It has been pivotal in unraveling contextual influences during reading such as parafoveal processing (**Kennedy & Pynte, 2005**), the impact of syntactic and semantic constraints (**Pynte et al., 2009**), and punctuation (**Pynte & Kennedy, 2007**).

While these studies show the intrinsic added value of corpora, the existing corpus data was scarce at that time, which resulted in several limitations. First, the narrative context was rather limited, as most eye tracking corpus data comprised sentence or paragraph reading. Second, and more importantly, only data of participants reading in their native language (L1) existed in the literature. As the knowledge of a second language (L2) is a global phenomenon (**Grosjean & Li, 2013**), bilingualism is an extremely relevant and important research domain in psycholinguistics, which could also benefit from dedicated eye tracking corpus data.

GECO

Recognizing the limitations in current datasets, we proposed the Ghent Eye tracking Corpus (GECO; **Cop et al., 2017**) to fill these gaps. By collecting eye movement data from monolinguals and bilinguals reading an entire novel, this pioneering bilingual corpus study aimed to bridge the divide between bilingual and monolingual reading research (see MECO, **Siegelman et al., 2022**, for a recent related project). Through this endeavor, we strive to provide profound insights into the reading process at multiple levels (e.g., word, sentence, or text), as well as language processing across diverse language backgrounds.

The study involved nineteen unbalanced Dutch-English bilinguals from Ghent University and fourteen English monolinguals from the University of Southampton. Participants were matched in age and education level, with an average age of 21.2 years for bilinguals and 21.8 years for monolinguals. All were enrolled in psychology programs, and none reported language or reading impairments. Bilinguals began learning their second language at the mean age of eleven years. The participants were subjected to several language proficiency tests, including a vocabulary test, a spelling test, a lexical decision task, and a self-reported test on language use and skill. The bilingual participants completed all tests both in L1 and L2. The two groups were equally proficient in L1, but the L2 English proficiency of the bilinguals was lower than the L1 English proficiency of the monolingual group.

The participants read the novel "The Mysterious Affair at Styles" by Agatha Christie (1920; in Dutch: "De zaak Styles"). The selection of this novel was based on its availability in various languages for potential future replication and its absence of copyright issues (it is included in the freely accessible Gutenberg collection). Further selection criteria which lead to the selection of this novel included the feasibility to read it within about 4 h, an examination of difficulty based on word frequency distribution, and an above-average reading ease based on readability metrics. Descriptives of the English and Dutch version of the novel are presented in **Table 1**.

The novel was presented on a computer screen in paragraphs up to 145 words. All participants read the novel in four separate sessions with a fixed number of chapters, spread over multiple days. Reading was self-paced, and to ensure that participants paid attention to the content, comprehension questions were administered after each chapter. Monolinguals completed the entire novel in English, bilinguals read half of the novel in Dutch (L1) and the other half in English (L2). The order of L1–L2 reading was counterbalanced, as such half of the bilinguals started reading in Dutch, the other half in English. Eye movements were registered monocularly at a sampling rate of 1000 Hz with a tower or desktop mounted Eyelink 1000 (SR Research, Canada).

The resulting data of GECO includes some 1.75 million datapoints. Per participant, both timed measures (e.g., first fixation duration and total reading time) and probabilistic measures (e.g., skipping during first-pass reading) are included for all words in the novel (for the

Table 1 Descriptives of the Dutch and the English version of "The mysterious case at Styles" as used in GECO.

	Dutch			*English*		
Number of words	59,716			54,364		
Number of word types	5575			5012		
Number of nouns	7987			7639		
Number of noun types	1777			1742		
Number of sentences	5190			5300		
	M	SD	Range	M	SD	Range
Number of words per sentence	11.64	8.86	[1–60]	10.64	8.20	[1–69]

Based on **Cop et al. (2017)**.

full range of measures, see the Appendices of Cop et al., 2017). Table 2 presents an example of average reading times and skipping probability for all groups. The complete corpus data is freely available via https://expsy.ugent.be/downloads/geco/.

The corpus approach offers distinct advantages in linguistic research. Firstly, it allows for assessing effects of continuous lexical variables, such as word frequency, across their entire range, enhancing precision compared to constrained settings. Secondly, large linguistic corpora enable researchers to explore multiple hypotheses without the need for new experiments, saving time and resources. Moreover, these datasets facilitate investigations at various levels of language processing, from word-level to semantic level. They support inquiries into diverse research questions regarding L1 and L2 reading.

While natural eye-tracking corpora offer valuable insights, they also come with limitations. Firstly, controlling confounding factors is challenging compared to controlled experimental settings. However, leveraging suitable metrics allows for covariate inclusion in statistical models, mitigating this issue. Secondly, due to the nature of the material, certain word characteristics or combinations may be underrepresented, such as combinations of extreme lengths (very short or long) with extreme word frequencies (very high or low). Generalizing results from these cases may be compromised due to the limited observations. Nonetheless, with over 5000 unique words in each language, meaningful results applicable to reading (in L1 and L2) can still be obtained. In the following section, we provide examples that show the wide range of applicability of GECO data.

Applications of GECO

In bilingual visual word recognition, Cop, Keuleers, et al. (2015) found that both bilinguals and monolinguals exhibit similar word frequency effects in their first language (L1), but bilinguals experience a larger frequency effect in their second language (L2), supporting the idea of an integrated mental lexicon influenced by language exposure. Dirix and Duyck (2017) extended findings on the age of acquisition (AoA) effect, revealing that words learned earlier in life are processed faster in both L1 and L2, with early L1 AoA also impacting L2 processing. Yaneva et al. (2017) demonstrated a processing advantage for multiword expressions in both native and non-native English speakers, though the advantage mainly appears in later processing stages. A final example by Dirix et al. (2017) showed that cross-lingual orthographic neighborhood affects both L1 and L2 reading, suggesting language-independent lexical access, Snell, Grainger, and Declerck (2018) found that embedded words (e.g., “arm” in “charm”) generally facilitate word recognition, leading to shorter viewing times and fewer fixations in Dutch and English readers. However, long, high-frequency embedded words showed inhibitory effects in Dutch readers, suggesting a dual mechanism of facilitation and inhibition. Yang, van den Bosch, and Frank (2022) highlighted the significance of sub-word and supra-word units in visual word processing, with cognitive units outperforming traditional word units in predicting fixation times.

Regarding sentence processing, Cop, Drieghe, and Duyck (2015) identified distinct L2 reading patterns, such as longer sentence reading times and increased fixations in comparison to L1 reading. Bilingual L1 reading however displays no significant deviations from monolingual reading behavior. Snell et al. (2023) found support for a prediction of their OB1-reader model, indicating that word order confusion arises from reliance on visual cues, particularly when words are of equal length.

In statistical learning, Snell and Theeuwes (2020) found that repeated exposure to specific multi-word structures enhances oculomotor control, with steeper learning curves for higher frequency structures. Finally, in cognitive modeling, Dotlačil (2018)successfully applied an Adaptive Control of Thought–Rational (ACT-R) model to simulate eye-tracking data, and Kun et al. (2023) introduced a dynamic approach to measure semantic similarity, which is capable of predicting fixation durations during reading, thus contributing to our understanding of language processing.

Table 2 The means, standard deviations and range of timed measures and skipping probability for monolingual, bilingual L1 and bilingual L2 reading.

	Monolingual (English)			*Bilingual L1 (Dutch)*			*Bilingual L2 (English)*		
	M	*SD*	*Range*	*M*	*SD*	*Range*	*M*	*SD*	*Range*
First fixation duration	214	70	101–502	209	65	101–467	222	74	101–536
Gaze duration	232	89	101–695	226	85	101–682	250	105	101–877
Total reading time	264	127	101–1060	256	117	101–852	296	194	101–978
Skipping probability	0.38	0.08	0.22–0.52	0.34	0.09	0.17–0.47	0.31	0.10	0.08–0.52

Based on Cop et al. (2017).

Expansions of GECO

To date, two expansions using the same source material as GECO have been published. First, GECO-CN (**Sui et al., 2023**) introduced a bilingual expansion of GECO, capturing natural reading behaviors of Chinese-English speakers. Particularly noteworthy is its potential to shed light on disparities between Eastern and Western languages, with completely different writing systems and orthographies, and to elucidate the effects of differing first languages on bilingual processing. GECO-CN allows for the validation or further exploration of the Chinese reading process, in particular the investigation of characteristics that are absent in alphabetic languages, such as stroke count of Chinese characters.

GECO-MT (**Colman et al., 2022**) included both a human and machine translated Dutch version of the novel as material. The rapid advancement of machine translation technology has significantly enhanced the quality of translated text in recent years. However, discernible distinctions persist between machine translations and human translations, particularly in the realm of more creative textual genres like literary works. Colman et al. focus on the end user of MT: the reader. Here, eye movement data was collected of participants reading half of the original human translated version of The Mysterious Affair at Styles, and half of a machine translated version. This allows for a comparison of reading behavior of individuals when reading machine translated compared to human translated text, hereby probing the extent to which machine translated text impacts the reading process.

Conclusions

In conclusion, the rich data and wide applicability of GECO emphasizes the vital role of eye tracking datasets of natural reading in advancing our comprehension of language processing and refining computational models of reading. The development of GECO represents a significant stride toward understanding the influence of language knowledge on reading processes, particularly in bilingual contexts.

References

Balota, D.A., Yap, M.J., Hutchison, K.A., Cortese, M.J., Kessler, B., Loftis, B., … Treiman, R. (2007). The English lexicon project. Behavior Research Methods, 39(3), 445–459. doi:10.3758/BF03193014.

Boston, M.F., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. Journal of Eye Movement Research, 2(1). doi:10.16910/jemr.2.1.1. Article 1.

Brysbaert, M., & Drieghe, D. (2024). The use of eye movement corpora in vocabulary research. Research Methods in Applied Linguistics, 3(1), 100093. doi:10.1016/j.rmal.2023.100093.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. Current Directions in Psychological Science, 27(1), 45–50. doi:10.1177/0963721417727521.

Chang, Y.-N., Monaghan, P., & Welbourne, S. (2019). A computational model of reading across development: Effects of literacy onset on language processing. Journal of Memory and Language, 108, 104025. doi:10.1016/j.jml.2019.05.003.

Clifton, C., Ferreira, F., Henderson, J.M., Inhoff, A.W., Liversedge, S.P., Reichle, E.D., & Schotter, E.R. (2016). Eye movements in reading and information processing: Keith Rayner's 40 year legacy. Journal of Memory and Language, 86, 1–19. doi:10.1016/j.jml.2015.07.004.

Colman, T., Fonteyne, M., Daems, J., Dirix, N., & Macken, L. (2022). GECO-MT: The Ghent eye-tracking corpus of machine translation. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., … Piperidis, S. (Eds.), Proceedings of the thirteenth language resources and evaluation conference (pp. 29–38). European Language Resources Association. https://aclanthology.org/2022.lrec-1.4.

Cop, U., Drieghe, D., & Duyck, W. (2015). Eye movement patterns in natural reading: A comparison of monolingual and bilingual reading of a novel. PLoS One, 10(8), e0134008. doi:10.1371/journal.pone.0134008.

Cop, U., Keuleers, E., Drieghe, D., & Duyck, W. (2015). Frequency effects in monolingual and bilingual natural reading. Psychonomic Bulletin & Review, 22(5), 1216–1234. doi:10.3758/s13423-015-0819-2.

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. Behavior Research Methods, 49(2), 602–615. doi:10.3758/s13428-016-0734-0.

Dijkstra, T., Wahl, A., Buytenhuijs, F., Halem, N.V., Al-Jibouri, Z., Korte, M.D., & Rekké, S. (2019). Multilink: A computational model for bilingual word recognition and word translation. Bilingualism: Language and Cognition, 22(4), 657–679. doi:10.1017/S1366728918000287.

Dirix, N., & Duyck, W. (2017). The first- and second-language age of acquisition effect in first- and second-language book reading. Journal of Memory and Language, 97, 103–120. doi:10.1016/j.jml.2017.07.012.

Dirix, N., Cop, U., Drieghe, D., & Duyck, W. (2017). Cross-lingual neighborhood effects in generalized lexical decision and natural reading. Journal of Experimental Psychology: Learning, Memory, and Cognition, 43(6), 887–915. doi:10.1037/xlm0000352.

Dotlačil, J. (2018). Building an ACT-R reader for eye-tracking corpus data. Topics in Cognitive Science, 10(1), 144–160. doi:10.1111/tops.12315.

Grosjean, F., & Li, P. (2013). The psycholinguistics of bilingualism. John Wiley & Sons.

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. Vision Research, 45(2), 153–168. doi:10.1016/j.visres.2004.07.037.

Kuperman, V., & Van Dyke, J.A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. Journal of Experimental Psychology. Human Perception and Performance, 39(3), 802–823. doi:10.1037/a0030859.

Pynte, J., & Kennedy, A. (2007). The influence of punctuation and word class on distributed processing in normal reading. Vision Research, 47(9), 1215–1227. doi:10.1016/j.visres.2006.12.006.

Pynte, J., New, B., & Kennedy, A. (2009). On-line syntactic and semantic influences in reading revisited. Journal of Eye Movement Research, 3(1). doi:10.16910/jemr.3.1.5. Article 1.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, 124(3), 372–422. doi:10.1037/0033-2909.124.3.372.

Reichle, E.D., Rayner, K., & Pollatsek, A. (2003). The E-Z reader model of eye-movement control in reading: Comparisons to other models. The Behavioral and Brain Sciences, 26(4), 445–476. doi:10.1017/s0140525x03000104. discussion 477-526.

Siegelman, N., Schroeder, S., Acartürk, C., Ahn, H.-D., Alexeeva, S., Amenta, S., … Kuperman, V. (2022). Expanding horizons of cross-linguistic research on reading: The Multilingual Eye-movement Corpus (MECO). Behavior Research Methods, 54(6), 2843–2863. doi:10.3758/s13428-021-01772-6.

Snell, J., & Theeuwes, J. (2020). A story about statistical learning in a story: Regularities impact eye movements during book reading. Journal of Memory and Language, 113, 104127. doi:10.1016/j.jml.2020.104127.

Snell, J., Grainger, J., & Declerck, M. (2018). A word on words in words: How do embedded words affect reading? Journal of Cognition, 1(1), 40. doi:10.5334/joc.45.

Snell, J., van Leipsig, S., Grainger, J., & Meeter, M. (2018). OB1-reader: A model of word recognition and eye movements in text reading. Psychological Review, 125(6), 969–984. doi:10.1037/rev0000119.

Snell, J., Mirault, J., Theeuwes, J., & Grainger, J. (2023). Readers use word length information to determine word order. Journal of Experimental Psychology: Human Perception and Performance, 49(6), 753–758. doi:10.1037/xhp0001107.

Sui, L., Dirix, N., Woumans, E., & Duyck, W. (2023). GECO-CN: Ghent Eye-tracking COrpus of sentence reading for Chinese-English bilinguals. Behavior Research Methods, 55(6), 2743–2763. doi:10.3758/s13428-022-01931-3.

Yaneva, V., Taslimipoor, S., Rohanian, O., & Ha, L.A. (2017). Cognitive processing of multiword expressions in native and non-native speakers of English: Evidence from gaze data. In Mitkov, R. (Ed.), Computational and corpus-based phraseology (pp. 363–379). Springer International Publishing. doi:10.1007/978-3-319-69805-2_26.

Yang, J., van den Bosch, A., & Frank, S.L. (2022). Unsupervised text segmentation predicts eye fixations during reading. Frontiers in Artificial Intelligence, 5. doi:10.3389/frai.2022.731615.