The Eye Movement, Reading, and Language Development (EMRLD) Corpus

Jeffrey Witzel (UT Arlington) & Naoko Witzel (UT Arlington)

This presentation will introduce a new database for investigations into first-language (L1) and second-language (L2) comprehension and reading processes – the Eye Movement, Reading, and Language Development (EMRLD) Corpus. In this database, reading performance in both L1 and L2 readers of English is assessed with measures of overall comprehension and reading speed as well as with recordings of readers' eye movements. These eye-movement measures are of particular interest because they allow for fine-grained analyses of real-time language comprehension and processes that underlie skilled reading [1]. The database also includes detailed information on the reading texts and on the readers' language and literacy backgrounds.

Several eye-tracking corpora have been developed for investigations into language comprehension and reading processes. These include databases that focus on L1 reading and, more recently, on both L1 and L2 reading [2-7]. The EMRLD Corpus complements these resources and extends the research questions that can be examined in important ways. It is the first corpus of its kind to systematically manipulate the complexity of the texts in the reading test. This will allow researchers to examine the influence of this factor on reading processes and its interaction with individual differences among readers in terms of their language and literacy backgrounds. Regarding these individual differences, this is also the first corpus of its type to focus on L2 English readers (i) who are studying in degree-granting programs at a US university and (ii) who come from a wide range of L1 backgrounds. This makes it possible to investigate the influence of language and literacy background on reading performance in a representative sample of L2 English learners at the highest levels of proficiency. Additionally, this is the first eye-tracking corpus to examine reading performance in a comparably large and diverse group of L1 English readers, including many simultaneous bi/multi-linguals. In these ways, this corpus offers a unique resource for researchers in linguistics, psychology, and education (among other fields) to investigate questions related to L1 and L2 comprehension and reading processes.

The database includes records for 95 L1 readers and 111 L2 readers. The L2 readers come from a wide range of language backgrounds, representing 23 L1s from eight language families. The reading texts are comprised of 18 passages (consisting of 2569 words) at three levels of text complexity. The presentation will focus on the reliability and validity of the corpus data for investigations into L1 and L2 comprehension and reading processes. It is first important to note that the L1 and L2 readers have comparably high accuracy rates on the comprehension questions for the reading passages (Figure 1), making it possible to draw clear inferences from other comparisons of their reading performance. Indeed, one of the most salient differences between these groups is that L2 participants read less efficiently than their L1 counterparts, as evidenced by their much longer per-word reading times on the passages (i.e., total reading time / # of words in the passage) (Figure 1). In terms of eye-movement measures related to this difference, compared to the L1 group, the L2 readers have a lower first-pass skipping rate, longer first fixation durations, longer gaze durations, more gaze fixations, a higher rereading rate, longer total times, and more total fixations (Figure 2). Crucially, there is very high reliability across these measures for the L1 and L2 readers at both the participant level (mean Spearman-Brown corrected reliability – L1: .98, L2: .99) and the word token level (mean Spearman-Brown corrected reliability – L1: .84, L2: .91). Regarding one of the unique contributions of this corpus, it is also important to evaluate its text complexity manipulation. As evidence of the efficacy of this manipulation, both L1 and L2 readers show longer per-word reading times with increasing text complexity (Figure 3). Finally, an analysis of L1 and L2 readers' log gaze durations as a function of log word frequency reveals a larger frequency effect for L2 readers (Figure 4). This replicates previous findings [8-10] and thus serves as another indicator of the quality of these data for examinations of L1/L2 processing.

The presentation will detail these and related findings, provide a complete account of the information provided in the corpus, and outline plans for its continued development. The corpus will also be made available to the research community so that it can contribute to other projects.

Figure 1. Mean comprehension accuracy rates and per-word reading times (in ms) for L1 and L2 English readers. Error bars indicate ±1 SE.

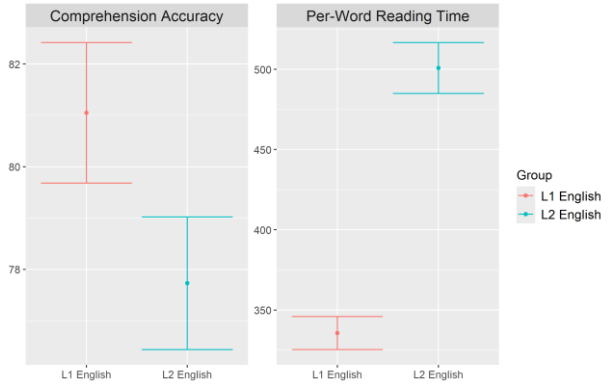


Figure 2. Means of eye-movement measures for the L1 and L2 English readers. Error bars indicate ±1 SE.

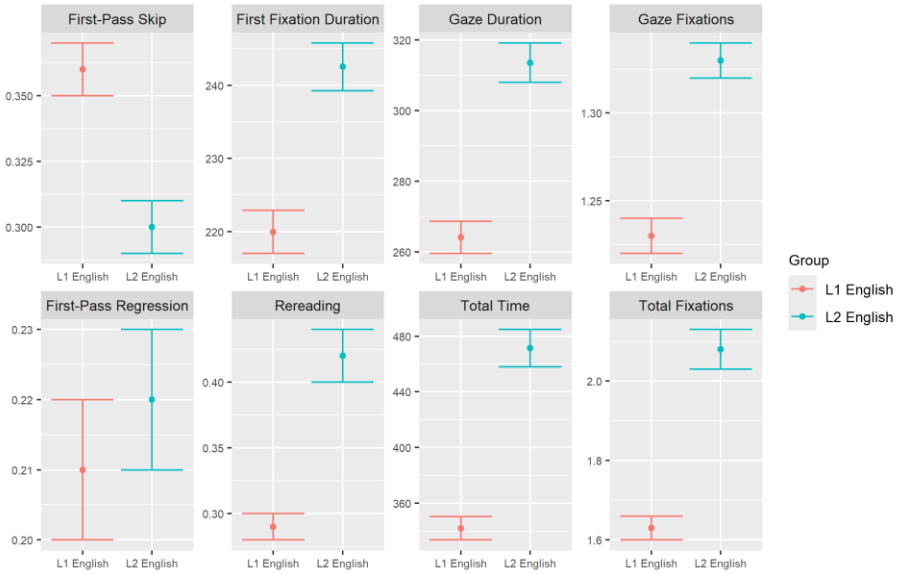


Figure 3. Means and 95% credible intervals for per-word reading time (in ms) at each text complexity level for L1 and L2 English readers.

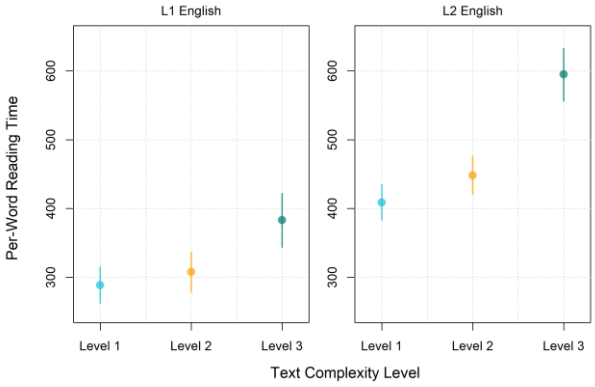
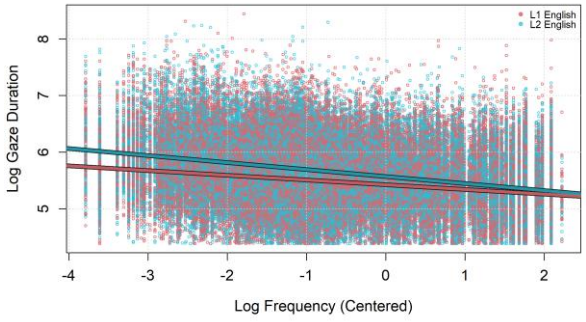


Figure 4. Log gaze durations plotted against centered log word frequency for L1 and L2 English readers, with regression lines for each group.



References

[1] Rayner, K. (2009). *QJEP, 62*, 1457-1506. **[2]** Berzak, Y., et al. (2022). *Open Mind: Discoveries in Cognitive Science, 6*, 41-50. **[3]** Cop, U., et al. (2017). *BRM, 49*, 602-615. **[4]** Kennedy, A. (2003). *The Dundee Corpus* [CD-ROM]. University of Dundee. **[5]** Kuperman, et al. (2023). *SSLA, 45*, 3-37. **[6]** Luke, S. G., & Christianson, K. (2018). *BRM, 50*, 826-833. **[7]** Siegelman, N., et al. (2022). *BRM, 54*, 2843-2863. **[8]** Cop, U., et al. (2015). *PBR, 22*, 1216-1234. **[9]** Duyck, W., et al. (2008). *PBR, 15*, 850-855. **[10]** Whitford, V., & Titone, D. (2012). *PBR, 19*, 73-80.