Familiarity and frequency disentangled: An eye-tracking corpus study with German texts

Sascha Wolfer, Sandra Hansen, & Lars Konieczny (University of Freiburg)
sascha@cognition.uni-freiburg.de

Eye-tracking during reading; Eye-tracking corpus study; Lexical processing; Frequency; Familiarity; Retrieval

Models of eye-movement control (e.g., E-Z-Reader, Reichle et al., 2003) make detailed predictions about the influence of lexical variables on the reading process. These predictions have previously been tested on large-scale eyetracking corpora like the Postdam sentence corpus (Kliegl, Nuthmann, & Engbert, 2006) and the Dundee corpus (Kennedy, Hill, & Pynte, 2003). One established predictor for reading times is word familiarity where highly familiar words show decreased reading times (Kennedy & Pynte, 2005). Gernsbacher (1984) had subjects rate words for their subjective, “experiential familiarity”. Here, familiarity is captured by the cumulated frequency of all words sharing word *n*’s length and initial trigram. However, familiarity is confounded with lexical frequency: Highly frequent words tend to be also highly familiar. Does familiarity contribute anything beyond word frequency?

In an analysis of lexical variables in reading German jurisdictional texts, press releases and newspaper articles, we applied a multi-residualization technique to assess this issue. Our eye-tracking corpus consists of gaze data from 80 participants on over 16,000 words. Word length, token frequency, familiarity, and the number of nearest neighbors as measured by a Levenshtein distance of 1 were extracted from the lexical database dlexDB (Heister et al., 2011) and treated as predictors in linear mixed-effects models (with participant and item as random factors).

For first fixation durations, first-pass reading times, regression path durations and total reading times, we found an effect of residual familiarity independent of the effects of word length and residual frequency that points in the opposite direction than the effect of raw familiarity. The fact that words with a high residual familiarity (with word length and frequency partialled out) are read longer points to an effect of lexical competitors. This effect is still reliable if the number of words with a Levenshtein distance of 1 to word *n* is included into the model, underpinning the relevance of residual familiarity capturing other sources of variance than lexical competitors. Obviously, the beginning and the overall shape (as captured by the length) of a word are important in lexical processing during reading. If there are many similarly shaped words with the same beginning, then lexical retrieval gets harder, which could be characterized as a fan effect (Anderson, 1974). The fan consists of all other words of equal length and the same beginning as word *n*. If the fan is smaller, fewer competitors interfere and word *n* gets activated faster. Hence, lexical retrieval is easier and faster.

In analyses of lag effects, we did not find an effect of familiarity of word *n*-1, whereas we found an interaction effect of the lexical frequencies of word *n* and *n*-1. The frequency of word *n* exerts a higher influence on first pass reading times when word *n*-1 is less frequent – a phenomenon captured by Henderson & Ferreira’s (1990) “foveal-load hypothesis”. This effect pattern is in line with a lexical retrieval explanation, because the early stages of lexical retrieval should already be completed when readers move on to the next word – therefore, no effect of residual familiarity of word *n*-1 is expected.

Analyses of lexical properties and their contribution to reading behavior emphasize the relevance of residual familiarity of word *n.* Even while reading natural texts, where semantics and pragmatics are constantly constraining the number of possible words which could be encountered next, the beginning and the overall shape of a word seem to influence reading times.

References

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology,* 6, 451-474.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. Journal of Experimental Psychology / General , 113 (2), 256-281.

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., et al. (2011). dlexDB - eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau,* 62(1), 10-20.

Henderson, J., & Ferreira, F. (1990). Effects of foveal processing difficulty on the perceptual span in reading: Implications for attention and eye movement control. *Journal of Experimental Psychology / Learning, Memory & Cognition,* 16, 417-429.

Kennedy, A., & Pynte, J. (2005). Parafoveal-on-foveal effects in normal reading. *Vision Research,* 45, 153-168.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin,* 124(3), 372-422.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z reader model of eye-movement control in reading: comparisons to other models. *Behavioral and Brain Sciences,* 26(4), 477-526.