Exploring a joint approach for analyzing reading and writing errors in Dutch

Wieke Harmsen; Roeland van Hout; Helmer Strik; Catia Cucchiarini
Radboud University Nijmegen, The Netherlands

In contrast to spoken language processing, written language processing requires skills that are not learned automatically such as reading (written language comprehension) and writing (written language production). Direct instruction and active practice are necessary to master the orthographic, phonetic and morphologic principles underlying these skills and automate the processing of written language.

Recent international research shows alarming decreases in reading skills in various countries and in particular in the Netherlands (Swart et al., 2023) where a decrease in children's spelling was also observed. These findings call for additional research into the development of children's reading and writing skills and for creative solutions that can halt these negative trends.

So far, research has focused on either reading or writing difficulties, while a large-scale combined approach has eluded researchers. This was probably due to the limited availability of large child reading and writing corpora and the fact that manual transcriptions and annotations are very time-consuming and costly. However, thanks to recent developments in language and speech technology, like grapheme-phoneme alignment, child word frequency lists, and high-quality automatic speech recognition (ASR) for Dutch, this kind of more comprehensive literacy research is now within reach. Together with the availability of the BasiScript corpus (Dutch texts and dictations written by children) (Tellings et al., 2018) and JASMIN corpus (Dutch texts read by children) (Cucchiarini et al., 2008), exploratory research into the relationship between spelling and reading errors in Dutch is now made possible.

In the current study we investigated which criteria a joint annotation scheme for Dutch reading and writing errors should comply with. In addition, we explored to what extent it is possible to automatically annotate reading and writing data with this annotation scheme. We discuss our findings and present avenues for future research.

Cucchiarini, C., van Hamme, H., van Herwijnen, O., & Smits, F.. 2006. JASMIN-CGN: Extension of the Spoken Dutch Corpus with Speech of Elderly People, Children and Non-natives in the Human-Machine Interaction Modality. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy.

Swart, N. M., Gubbels, J., in 't Zandt, M., Wolbers, M. H. J., & Segers, E. (2023). PIRLS-2021: Trends in leesprestaties, leesattitude en leesgedrag van tienjarigen uit Nederland. Expertisecentrum Nederlands

Tellings, A., Oostdijk, N., Monster, I., Grootjen, F., & van den Bosch, A. (2018). BasiScript: : A corpus of contemporary Dutch texts written by primary school children. International Journal of Corpus Linguistics, 23 (4), 494–508. doi: https://doi.org/10.1075/ijcl.17086.tel