Evaluating Voices of Groningen

An interactive web-based approach to collecting Low Saxon dialect data

Raoul Buurke, Nanna Hilton, Martijn Wieling
University of Groningen

The *Stemmen van Grunnen* ('Voices of Groningen') web application (available at https://woordwaark.nl/stemmen) is aimed at collecting dialect data about the Low Saxon dialects within the Netherlands. It is based on *Stimmen van Fryslân* ('Voices of Friesland'; see Hilton, 2019) and similar applications (Leemann et al., 2016) in which the dialect of a speaker is located geographically on the basis of their selected dialectal variants for several words. Our application moreover asks the participant to record their own variants, which allows for investigating finer-grained dialect differences than when only selected variants are used. Advantageously, the acoustic recordings should no longer need manual transcription, as – in theory – the chosen provided transcription can be used.

Each participant is first asked to select the dialectal variant closest to their pronunciation for 10 Dutch words. The available variants were determined on the basis of the Goeman-Taeldeman-Van Reenen project (GTRP; Taeldeman & Goeman, 1996). Each variant is presented visually (in an intuitive spelling and IPA) and acoustically (if desired), after which participants are asked to record their own dialectal variant. Finally, the predicted geographical location of the speaker's dialect is shown on the basis of a GTRP-based decision tree.

Currently, over 1900 speakers have participated (i.e., approximately 19000 recordings). Before 2021, most data were collected from elderly speakers in the province of Groningen. Stemmen was included in a large-scale questionnaire covering a larger area in 2021, which doubled the amount of data.

The recordings of 377 words (10% of the available data at the time) were manually transcribed for evaluation purposes. Subsequently, we calculated the Levenshtein distance (Levenshtein, 1966) between these transcriptions and the transcriptions of the selected variant. The Levenshtein distance is a popular approach in dialectometry to quantify the difference between pronunciations (e.g., Kessler, 1995, Heeringa, 2004, and Wieling, 2012). The average Levenshtein distance was 0.5 (SD = 0.4), whereas the averaged normalized (over alignment length) Levenshtein distance was 0.09 (SD = 0.06), indicating a generally small difference between the actual pronunciation and the selected variant.

To assess whether participants chose the variant closest to their pronunciation, we ranked each Levenshtein distance between the option per word per participant and their pronunciation. We then normalized the ranks between 0 and 1 (with 0 representing the best possible choices and 1 the worst ones). In 16% of all cases, participants produced a form not present on the list. In a minority of these cases (28%) participants selected a non-optimal variant. In the 84% of cases when the pronounced variant was on the list only 10% of the time a non-optimal variant was selected. The feasibility of our approach is reflected by the close-to-optimal average normalized rank of 0.04 (SD = 0.05).

In sum, our approach with Stemmen van Grunnen seems suitable to obtain dialectal recordings together with (automatic) transcriptions. This opens the door to use both transcription-based dialectometric techniques, but also acoustic-based techniques to quantify pronunciation differences.

References

Heeringa, W. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. PhD thesis, University of Groningen.

Hilton, N. (2019). Smartphone sociolinguistics in minority language areas: 'Stimmen'. Published: Paper presented at International Conference on Language Variation in Europe (ICLaVE) 10, Leeuwarden.

Kessler, B. (1995). Computational dialectology in Irish Gaelic. In Abney, S. P. & Hinrichs, E. W. (Eds.), *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*, (pp. 60–66). Morgan Kaufmann Publishers Inc.

Leemann, A., Kolly, M.-J., Purves, R., Britain, D., & Glaser, E. (2016). Crowdsourcing language change with smartphone applications. *PLOS ONE, 11*(1), e0143060.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Taeldeman, J. & Goeman, A. (1996). Fonologie en morfologie van de Nederlandse dialecten: Een nieuwe materiaalverzameling en twee nieuwe atlasprojecten. *Taal en Tongval, 48*, 38–59.

Wieling, M. (2012). *A quantitative approach to social and geographical dialect variation*. PhD thesis, University of Groningen.