Speak for Yersel: crowdsourcing Scots in the 21st century

The digital age has revolutionised how we collect and analyse dialect data. The pen and paper methodologies of the 19th century (e.g. Ellis 1889) were replaced by large-scale, computer-aided corpora in the 20th century (e.g. various ICE corpora http://ice-corpora.net/ice/index.html), and further advances in technologies in the 21st century has seen a move towards crowdsourcing (e.g. Leemann et al 2018), the practice of obtaining information from a large number of speakers via online means.

In this talk we report on a new digital resource, *Speak for Yersel*, which crowdsources the dialects of Scots used throughout Scotland across lexical (1), phonetic (2), and morphosyntactic forms (3), and attitudes towards these.

1. That *quine* got the job.
2. I was [u:]t and ab[u:]t the day
3. *Gonnae you* lift that for me?

We detail the issues that arose in constructing *Speak for Yersel*, focussing on the central tension between resource accessibility and the sociolinguistic complexity that abounds in the Scots spoken across Scotland, and in language variation and change more generally. Specifically, how such pressures impact on the content to include, and how it is best presented in order to tap sociolinguistic norms.

We then turn to an analysis of the data collected in *Speak for Yersel*. The resource launched in October 2022, with dissemination largely through Twitter. By the end of the first week it had logged c6000 users and c500,000 data points. Fine-grained statistical analysis of the production data (c400,000 data points) across linguistic and social constraints suggest that variation is successfully accessed through online means with some levels of language, but not so successfully with others. In particular, the (lack of) patterns of use point to a general cline in ability of speakers to report faithfully on what they do or do not say: lexical > phonological > phonetic > morphosyntactic, with further variations within these.

We discuss what these results reveal about a speaker's ability to tap into their own language use in the context of dialect data (e.g. Labov 1996), and how they can inform more generally on attempts to capture complex sociolinguistic patterns through necessarily simplified virtual means.

References

Ellis. A. (1889). On Early English Pronunciation: Part V Truebner and Co, London.

Labov, W. (1996) "When intuitions fail" in McNair, L. et al. (eds.) Papers from the Parasession on Theory and Data in Linguistics: CLS 32 Chicago: Chicago Linguistic Society pp. 77-106.

Leemann, A., Kolly, M-J. and Britain, D. (2018) The English Dialects App: the creation of a crowdsourced dialect corpus. Ampersand, 5. pp. 1-17.