Integrated Speech Corpus ANalysis – ISCAN
A new tool for large-scale, cross-corpus, sociolinguistic analysis

Jane Stuart-Smith (University of Glasgow), Morgan Sonderegger (McGill University), Michael McAuliffe (McGill University)

Our workshop will introduce variationist sociolinguists, especially those working with spoken language, to a new software system, *Integrated Speech Corpus ANalysis (ISCAN)*, which enables fast, high-quality, large-scale automated acoustic phonetic analysis across multiple spoken corpora of diverse formats, simultaneously[1]. *ISCAN* features an easy-to-use graphical interface (in a web browser), and can also be used via a Python package for more complex analyses by technically-skilled users. We will describe how ISCAN works, and then give participants the opportunity to use the software themselves to carry out analyses of vowels and sibilants, examining commonly-measured variables (formants, COG, speech rate, etc.) across publicly-available datasets. Participants will gain sufficient experience with *ISCAN* to understand how to apply the system to their own data, and perhaps to do so on-site.

ISCAN is being developed as an open, freely-accessible, software system, by the *SPeech Analysis across Dialects of English(SPADE)* project, with colleagues from North America and the UK, specifically, Jane Stuart-Smith and Joe Fruehwald (UK), Morgan Sonderegger and Michael McAuliffe (Canada), Jeff Mielke, Erik Thomas, Robin Dodsworth, Tyler Kendall and Paul Fyfe (US) (2017-2020: https://spade.glasgow.ac.uk/). SPADE is primarily a methodological project, which is developing software tools to (a) assemble and import spoken language corpora of diverse formats, (b) enrich these corpora with additional linguistic information, e.g. lexical frequency, parts of speech, syllabic information, (c) carry out automated speech processing across one or more speech corpora to generate high-quality acoustic phonetic measures for speech segments, along with key durational and prosodic measures; and (d) allow users to query and extract measures for segments for subsequent analysis, using an accessible user interface, the *ISCAN* system which we will introduce at the workshop. Our linguistic remit is currently some varieties of English, but *ISCAN* can be used for any linguistic variety, provided that the sound file has an accompanying segmentation file (e.g. from a forced alignment using FAVE, MFA, or LABBCAT).

We are currently working with a number of sociolinguists and phoneticians in the UK and North America, who are sharing existing speech datasets, for the development of the software. Analyses so far, presented at LabPhon 2018, analyse data from about 10 English dialects (in the UK, US, and Canada). We are very keen that *ISCAN* will be as useful as possible for a wide range of users, from less to more experienced, and so feedback from NWAV47 participants will be extremely valuable. A key project goal is to build out more complex versions of the software by incorporating feedback from users about what does and does not work in the interface, and altering the software to accommodate users' request. Giving this workshop comes at a key stage in our project (start of year 2), which will lead to software which can be more responsive to the needs of sociolinguists.

[1] ISCAN is an updated version of Polyglot-SCT, see: McAuliffe, Michael, Elias Stengel-Eskin, Michaela Socolof, and Morgan Sonderegger. "Polyglot and Speech Corpus Tools: a system for representing, integrating, and querying speech corpora." Proc. Interspeech 2017 (2017): 3887-3891. See: https://github.com/MontrealCorpusTools/PolyglotDB

SPADE is working with publicly available datasets, and private datasets, managed by team members, and members of the sociolinguistic and phonetic community. A key consideration for working with private datasets is the appropriate and ethical management of user access, to ensure that those who do not have permission to hear or identify speech recordings, are not able to do this. Only publicly-available datasets will be used for the workshop. One goal of SPADE is to enable private speech datasets to be analysed for their acoustic speech features, without the analysts needing to listen to the speech recordings, or see the transcripts, what we are calling, 'ethically non-invasive speech corpus analysis'. To this end, we are devising a set of inspection interfaces which allow users to see different 'views' of parts of the speech corpora, for example, to check likely erroneous items/measures, but which also work in conjunction with user permissions for access to the corpora, or parts of the corpora. A beta version of these inspection interfaces should be available by the workshop, which would be a valuable opportunity for users to provide feedback.