ICE-Scotland – a phonologically annotated corpus

Ulrike Gut[1], Ole Schützler[2], Zeyu Li[1] & Jenny Herzky[2]

[1]University of Münster, [2]University of Bamberg

This paper presents ICE-Scotland, the first ICE corpus to contain time-aligned phonemic transcriptions. The corpus is still under construction (see Schützler et al. 2017), but will eventually contain 1 million words (400,000 words of written and 600,000 words of spoken language in the text categories specified in the ICE guidelines) of Scottish English produced by educated speakers from all over Scotland, who were born and brought up there. We will first report on the corpus compilation process, which includes three types of annotation:

(i) the time-aligned orthographic transcriptions divided into intonational phrases carried out using ELAN (https://tla.mpi.nl/tools/tla-tools/elan/download/) and Pacx (http://pacx.sourceforge.net/);

(ii) the automatic time-aligned phonemic transcriptions created by WebMAUS (Schiel 2004) and their manual correction in Praat;

(iii) the automatic addition and manual correction of POS-tags created with the *ClawsAnt* tagger using the CLAWS7 tagset.

This will include a critical evaluation of our decisions at every stage of the process. Further, we expressly invite discussions of the feasibility and usefulness of combining phonemic and grammatical (POS-based) annotation in a single format, which is possible but technically demanding and currently not yet implemented in our corpus.

In the second part of the talk we will show how ICE-Scotland can be used to explore (variation in) phonological features of Scottish English by presenting a case study of the NURSE vowel. Scottish Standard English is traditionally described as not having undergone the NURSE merger, a process which merged the Middle English vowels /ɛ, ɪ, ʊ/ into the vowel /ɜ/ in prerhotic positions but to have kept a three-way distinction in these contexts instead (e.g. in the words *bird*, *earth* and *nurse*). However, the gradual loss of this contrast has been observed in some varieties of Scottish English (e.g. Lawson et al. 2013; cf. Stuart-Smith 2008). Our analysis shows that in purely acoustic terms the vowels in *fir*, *fern* and *fur* are not merged in formal Scottish English but have a distinct F1 and F2. However, the prerhotic items are distinct from the reference categories KIT, DRESS and STRUT in being more centralised, and in some genres *fir* and *fern* are more strongly drawn towards the centre of the vowel space (and each other) than *fur*, with orthography and the realisation of the following /r/ having a significant effect.

Apart from the results that we present, we will illustrate our work routines (employing Praat, Excel, and R in a hybrid approach) and highlight issues and problems that remain despite the careful annotation scheme that we followed.

References

Lawson, Eleanor, James M. Scobbie & Jane Stuart-Smith. 2013. Bunched /r/ promotes vowel merger to schwar: An ultrasound tongue imaging study of Scottish sociophonetic variation. *Journal of Phonetics* 41(3-4): 198–210.

Schiel, Florian. 2004. MAUS goes iterative. In Maria Teresa Lino et al. (eds.), *Proceedings of the 4th International Conference on Language Resources and Evaluation, Lissabon.* Paris: European Language Resources Association. 1015–1018.

Schützler, Ole, Ulrike Gut & Robert Fuchs. 2017. New perspectives on Scottish Standard English. Introducing the Scottish component of the International Corpus of English. In Sylvie Hancil & Joan Beal (eds.), *Perspectives on Northern Englishes*. Berlin: Mouton de Gruyter. 273–301

Stuart-Smith, Jane. 2008. Scottish English: Phonology. In: Bernd Kortmann & Clive Upton (eds.), *Varieties of English*. Vol. 1: *The British Isles*. Berlin: Mouton de Gruyter. 48–70.