Speech Corpora for Language Learning

Guy ASTON & Daniele RODI
Advanced School of Modern Languages
for Interpreters and Translators (SSLMIT)
University of Bologna
Forl ì, Italy
christopherguy.aston@gmail.com; rodi.daniele@gmail.com

A brief glance at the CEFR reminds us that standard language learning objectives are largely formulated in terms of oral skills. But the corpora used in language teaching and learning are mainly composed of written texts, with only limited quantities of spoken transcripts. The quantitative imbalance is also a qualitative one, since transcripts do not provide teachers and learners with authentic access to speech as an oral phenomenon (Braun, 2009), with its pronunciation, intonation, segmentation and timing, and in consequence most work with spoken corpus data focuses on lexicogrammatical features. However, some corpus interrogation software (e.g. Scott's *Wordsmith* 5 and 6, Szakos & Glavitsch's *Speech Concordancer*) now allows access to the original audio as well as to transcripts in concordancing, provided the audio and the transcript are aligned. But like transcription, alignment requires considerable manual effort, making it prohibitive to produce other than very small aligned corpora – unless, that is, one can make use of pre-transcribed and pre-aligned material.

This paper describes an attempt to produce a larger aligned speech corpus by exploiting an obvious source of pre-transcribed and pre-aligned material, namely subtitled videos on the web. While much such material (e.g. films and television programmes) is of course restrictively copyrighted, some sources do provide more freely available data. One is the non-profit TED (Ideas worth spreading) site (www.ted.com), which distributes video and audio versions of talks and discussions on a wide variety of topics under the Creative Commons licence, often with eminent speakers (from Steve Jobs/Bill Gates to David Cameron/Michelle Obama). Many of these videos have crowd-sourced subtitles, providing reasonably accurate if broad transcriptions (repetitions, false starts, fillers and pauses are generally omitted). So far we have put together a corpus of 500 TED talks, containing over 1 million words of transcript and over 100,000 segments aligned with the corresponding audio files. The alignment information is included in the transcript files, which we have marked up in a pseudo-XML format designed to be compatible with *Wordsmith* (which unfortunately remains not fully XML-aware). In the file headers we have added categorizations of the talks by speaker dialect, sex, and topic, allowing searches to be restricted using these parameters where desired.

We are currently trialling this corpus in EFL lessons on our postgraduate degree course for trainee conference interpreters. The talks - mainly fairly spontaneous monologue - seem perceived as interesting and relevant to learners' listening and speaking objectives. One of our major teaching emphases is on an arguably key feature of fluent listening and speaking, that of spoken phraseology (Lin, 2010; Lin & Adolphs, 2009), focusing not simply on collocation/colligation but also on chunking and intonation, stress placement, articulation rate and reduced pronunciation. The paper will exemplify use of the corpus in these areas by and with students, and discuss the potentials and limits of similar corpora as learning and teaching resources in the development of oral skills.

References

Braun, S. (2009). 'Corpus-enhanced language learning and teaching.' Available from: http://www.es-courseportal.uni-tuebingen.de/backbone/moodle/file.php/1/reports/_7_bb_deliverable3-1_corpus-enhanced_LLT_part_1.pdf.

Lin, P. (2010). 'The phonology of formulaic sequences: a review'. In Wood D. (ed.). *Perspectives on Formulaic Language*. London: Continuum. 174-193.

Lin, P. & Adolphs S. (2009). 'Sound evidence: phraseological units in spoken corpora'. In Barfield A. & Gyllstad H. (eds). *Collocating in Another Language: Multiple Interpretations*. Basingstoke: Palgrave Macmillan. 34-48.