SALTMIL Special Session at Interspeech 2007

August 2007, Antwerp, Belgium

Report by Briony Williams

The "Interspeech" conference is an annual global conference in speech science and technology, which takes place alternately in Europe and in other continents. In 2007 it was held in Antwerp, Belgium, in the Flanders Congress and Concert Centre. More details can be found at http://www.interspeech2007.org

This Special Session was a poster session, chaired by Briony Williams, on the afternoon of Wednesday August 29th. The timetable is at http://www.interspeech2007.org/Technical/less_resourced_languages.php . Fifteen posters were presented, together with three demo papers (which did not feature in the conference Proceedings volume). Many different languages were represented, including Hungarian, Arabic dialects, Serbian, Gikuyu, Icelandic, Amharic, Thai, and Formosan Austronesian languages. In addition, several of the papers concerned tools that would be applicable to any language under investigation, such as the following:

- TurboAnnotate: open-source software for carrying out either manual or semi-automatic annotation of data.
- Web-based creation of speech resources by hand, using a client-server model rather than locally-installed software.
- VoiceTRAN machine translation system: this is a statistical machine translation system, developed using freely available tools and language resources. It was tested on English and Slovenian data, and performed better than a rule-based system.
- MuLAS: A framework for automatically building multi-tier corpora. The Multi-Level Alignment System (MuLAS) has been developed as a tool for building multi-level speech corpora either partially or fully automatically. It has been used for the speedy development of multi-tier corpora for speech synthesis in several languages.
- LEXUS: A flexible web-based lexicon creation tool, targeted at linguists working with less-resourced languages. It allows lexica to be created compliant with the proposed ISO LMF standard, and uses the proposed concept naming conventions from the ISO data categories. It also allows the linguist to add audio, video and still images to the lexicon. It is available free of charge.
- Morfessor and VariKN: Machine learning tools for speech and language technology. These are two open-source tools for unsupervised natual language modelling. Morfessor segments words automatically into morpheme-sized units without using rule-based morphological analysis. VariKN is a toolkit for training language models.

- ELAN: This is a multimedia annotation tool available free of charge for multiple platforms. It is widely used.
- SpeechIndexer: This software has been developed for annotating and retrieving spoken corpus data for endangered Formosan languages.
- A laser-beam reflection method for playback of damaged or fragile wax recording sylinders. Many wax cylinders are too fragile to be played in the original manner, but this new laser-beam methods can access the recorded data without the need to compromise the physical media.