Short Report on the CMLD (Computational Methods for Endangered Languages) conference

Joseph J Mariani

Short Report on the CMLD (Computational Methods for Endangered Languages) conference
(ENS, Paris, 1-2 February 2018)

J. Mariani

I participated in the CMLD (Computational Methods for Endangered Languages) conference last week at ENS in Paris. Laurent Besacier was also attending. Here is a short report.

- About fifty participants, several were also present at the workshop on Uralic languages in Saint Petersburg a year ago
- Proposal to use facial recognition for metadata on speakers' names (Niko Partanen, one of the organizers).
- Joachim Nivre presented the international effort on "Universal Dependencies (Treebanks)". Already 60 languages. I ask him what he plans to do for the remaining 6840... He says it's a funding problem. He agrees that they should be treated by family of languages to save money.
- Presentation by Jargal Badagarov (Mongolia: bjargal@mail.ru) on the Buriat language spoken in Mongolia, Russia and China. They have a collaborative network. Together with Laurent, we proposed him to establish links between SIGUL and this network. He uses the motto : "Some data is better than no data". I told him it's funny that we are used to say in our community "There is no better data than more data." !
- Laurent presented the Bulb project. Lot of interest.
- Timofey Arkhangelski on the Beserman Udmurt documentation project. I asked him if all recordings were translated into Russian just as in the Bulb case with French in the case of African languages, as all his speakers were bilingual. He says their group of linguists don't need it as they speak the language. But what about others ? From this discussion, I thought that it would be nice to ask all project on spoken language documentation to provide translation into a reasonably-well resourced language (English, French, Russian, Spanish, Portuguese, Mandarin), which benefit from Language Technologies, as it would provide a link between the signal in the source language and the meaning in the target language. This would enlarge the community of people working on those source languages, including computational scientists, and facilitate the resource documentation and use. I believe it could be a mission for SIGUL to ensure that it is done and exchanged.

- Francis Tyers on Speech Synthesis for endangered languages (Turkic) : "An endandered language will make progress if it can use electronic language technologies"
- 5 posters on the first day. Pierre Magistry (LIMSI) on Deep Learning on small data for POS tagging of LRL. Laurent on large scale phonetics: automatic analysis of vowel length in Wolof. It allows linguists to study much larger quantity of data than if purely manual. Alexis Michaud/Graham Neubig on phonemic transcription of low-resourced tonal language. Collaboration computer scientists-linguists. They reach 15% error rate on phonemes!

- 8 posters on second day. construction of Kabylian language database with Nooj. Berber language spoken in Algeria. They use latin alphabet, while in Morocco IRCAM uses Amazich alphabet... Also use of Nooj for studying the Rromani language and its 4 main dialects (Masako Watabe (Univ. Sorbonne). Problem for typing characters in Rromani dialects. I told her that Khalid chairs an ISO group on HMI taking care of keyboards. Request from Maximiliano Duran (duran_maximiliano@yahoo.fr) to provide a speech recognizer for transcribing (either words or phonemes) a database in Quechua.

Contact with Alexandre Lissy (Mozilla, Paris: alissy@mozilla.com). Looks for collaboration with laboratory for LRL TTS and ASR (Common Voice): choice of appropriate training data.