Summary of the discussion on "Less resourced languages and Language technology. Short- and medium-term objectives"

Valletta, Malta, 23 may 2010

After a brief presentation by Mikel Forcada (see slides), an interesting discussion took place.

One of the problems that was underlined is the difficulties in convincing politicians to fund the creation of language resources (LR) for less-resourced languages (LRL). Per Langgård suggested that it would be necessary to build a scheme to assist developers to have success in that endeavour; Khalid Choukri said that even for large European languages it was also difficult to convince European Union politicians to fund R&D in the field, and that we needed to give politicians a larger picture and something they can sell to the media. Along the same lines, Igor Leturia mentioned that we should convince politicians that we do not only do research but that we produce products that politicians can see.

Trond Trosterud proposed that the threshold to access LRs or tools should be as low as possible, and that friendly ways to disseminate should be put in place (one-click, grab-and-go interfaces). Benoît Sagot underlined the importance of finding ways to reduce the costs of developing LRs. Mikel Forcada suggested that the Recaptcha! idea (a test used in webpages to test that they are being accessed by a human) could be adapted to generate or test lexical language resources collectively. In connection with this, it was suggested that linguists should be given status and credit that is equal to the one computer scientists or engineers get, and that where tested this happened to work well.

A participant mentioned that perhaps it was better to create a single, well annotated and well evaluated resource; in the case of Arabic, the Qur'an could be collaboratively annotated with different layers of linguistic annotation in the hopes that that knowledge could be then extrapolated to the whole language.

Khalid Choukri mentioned that one of the problems with scientific literature dealing with language resources for LRLs is that many research groups created resources that were already available, and that ignorance about other research in the field was indeed a problem. He advised us to always make sure that our research is available to everyone, with an open license, and that the best practices in the field should always be applied when creating new resources. He also mentioned the importance of talking to publishers of printed books to build text resources, by offering them a business model that may be mutually beneficial.

Benoît Sagot insisted that LRs should be made as free as possible so that they are widely used, and mentioned that catalogues of language resources such as ELRA or LDC contained many non-free resources. Khalid Choukri responded that ELRA pricing and conditions just reflect the will of the authors of those resources, and that authors can obviously place free resources in the ELRA catalog. Mikel Forcada mentioned that free/open-source LRs offered an excellent way to ensure reproducibility in LR research, which is crucial to R&D, and also to be aware that R&D

grows in the directions that the society, through their funding agencies, decide, and that it is sometimes hard to convince decision-makers about the importance of less-resourced languages.

Kepa Sarasola mentioned the example of Basque: orthographical standardization was adopted in 1968, and spelling checkers developed in the 90s have been one of the most powerful ways of disseminating and promoting the standard orthography of unified Basque.

At 13.55, Mikel Forcada closed the panel session by thanking the audience for their participation in the panel and in the whole workshop.