Computel partnerships in practice: GiellaLT

GiellaLT

February 23, 2024

1 Abstract

In language technology small projects and experiments keep popping up with impressive results when looking at their statistics and evaluations. The question that stays unanswered in many of these projects, however, is how any of these greatly working tools will *benefit an actual community of language users*, how or if at all they will be implemented and made accessible to them and if they even are needed and asked for.

In practise an *active dialogue with a user community* can only be established when ongoing support and maintenance is available. This typically requires a partnership between language communities and computational linguists and often also governments support or other ways of funding. Divvun/Giellatekno is an example of an ongoing partnership for more than 20 years. This partnership came into being through an initiative of individuals that had a foot in the door with the Norwegian Sámi Parliament and Uit - The Arctic University of Norway. The group includes people from the community and other experts in language technologies with expertise in computational linguistics and the target languages.

One of the challenges for such a project is that *one needs dedicated people from the language community* to work and collaborate with. A starting point is ideally a natural arena (like in this case a university) where the experts meet on the same level, both Sámi and non-Sámi, and start working together completing each others expertise. For meaningful collaborations, positions need to be attractive to people, they cannot be below standards or potential applicants will choose other things. This work also requires a certain flexibility regarding work places to make sure that potential co-workers can be within their communities or spend a certain amount of time there while working. Contacts to schools, media, recruiting etc. often go via insiders, that means co-workers who know somebody within each field or sub-community. "Knowing someone" is the real key factor here.

Work on Sámi language technology started around the year 2000, where the university and later the Norwegian Sámi Parliament started to work on Sámi language technology. The Parliament had practical needs to fulfil, the goal of the university was to combine basic research with working programs. During the subsequent two decades the groups have joined at the university and now contains around 10 researchers. The rationale for continuous and increasing funding has all the time been the groups' ability to combine research with developing relevant tools for various indigenous and minority language communities.

With the number of languages as well as number of the technological tools and applications growing, a need arose to separate language-dependent and language-independent code, and to reuse as much as possible in order to get maximal benefit from the *scarcity of resources* (be it funding or staff) small speaker communities always experience. This means that we have gone from only having handful of languages and spelling checker to dozens of languages and grammar checking and correction and dictionaries and speech recognition and synthesis and so on and so forth and instead of re-starting linguistic work from zero we can work very incrementally.

What makes this work stand out as **unique also on a global scale** is the combination of two factors:

1. The work on Sámi language technology has been funded for a long period, and funding continues;

2. A substantial part of the funding has been spent on making the infrastructure portable to new languages, i.e., the funding allocated to meet the needs of Norway's minority language policy can be utilised by everyone.

To date, *more than 50 indigenous or minority language communities* have benefited or are about to benefit from this. The purpose of this presentation is to explain the collaborations that we have to make this all work well.

In the recent years of language technology the *text and speech corpus resources* have become more precious, and also the intellectual property rights as well as responsible usage of data has become more important topics. Because of this it is very important for us that we collect all the language data with cooperation of the language users and communities, and that licensing and usage rights are handled correctly both from legal point of view but also ethically and respecting the rights of language users.