Nooj conference 2014

Title :

Multi-dialectal and multi-linguistic e-dictionary and spellchecker for Rromani

--Nooj module for Rromani integrated in Smallcodes platform

Co-authors :

Carlo Zoli (Smallcodes S.R.L.) <carlo.zoli@smallcodes.com>

Masako Watabe (INALCO) <masakowatabe@free.fr>

Abstract :

Living in a hyper-connected world means that most of the inputs we receive daily are mainly written. Therefore only those inputs with the right features, such as a good visibility, the right patterns and an understandable language are winners in the written media. It is clear that, if we want to save and preserve minority languages, we must necessarily let them have access to the tools and resources of the same technological level as those of “bigger” languages.

Smallcodes (http://www.smallcodes.com/home.page?country=eng) is a commercial firm founded with the aim of creating a single and integrated platform for language technology dedicated to minority languages. The big novelty in the design of this instrument is that they put all the tools together in one single highly interoperable box. Till now they have contributed to several projects for minority languages such as Sardinian, Occitan and so on.

Smallcodes currently works for the "R.E.D.-RROM (Restoring the European Dimension of Rromani Language and Culture)” project (http://red-rrom.eu/home.page#). This is a self-learning system about Rromani language and culture. Within this project, Smallcodes has imported the first Rromani dictionary with multi-linguistic translation (Marcel Courthiade, “Morri angluni rromane ćhibǎqi evroputni lavustik”, 2009) to their electronic platform.

The paper version of the dictionary includes the four basic dialectal variants of Rromani (O and E superdialects, each without, or with, phonetic mutation) and translation in 10 languages (Croatian, English, French, German, Greek, Hungarian, Romanian, Slovak, Spanish and Ukrainian). These dialects share most of vocabulary and grammar, and mutual understanding is completely possible. There is no need to create a dictionary and a grammar separately for each dialect. There is no totally unified unique standard Rromani nor priority dialect representing a unique standard, so we treat all four dialects even. Rromani speakers are scattered all around the world, mainly in Europe but also across the Atlantic. If they had a solid electronic source in their own and only language, not only the distance communication, but also the sense of belonging to Rromani culture would be encouraged.

Nooj module for Rromani (Masako Watabe, 2011) contains a small dictionary with very few entries but each of which is morphologically expanded through a paradigm. The list of nouns, verbs and adjectives is quite complete and it includes exceptions and dialectal variants. It would be worthwhile to integrate Nooj morphology in Smallcodes platform in order to create a spellchecking systems of all Rromani dialects.

Smallcodes’ list of entries would be matched with the list of words with corresponding paradigms pre-elaborated in Nooj. In fact, Smallcodes platform, which already contains a module for spellchecking creation through morphological expansion, can acquire rules (i.e. paradigms) from external sources. The spellchecker would be then integrated in the e-dictionary and the two modules would constitute a first and crucial resource for all those Rrom students who want to acquire their language and use it proficiently in its written form.

In addition, once these two modules for Rromani have been completed in Smallcodes platform, we could re-export them in Nooj to develop further grammars, especially concerning syntax (currently Smallcodes platform does not contain syntax). This would allow to annotate texts in a more appropriate way, to develop an archive from various corpus written in different dialects and so on. Then Smallcodes platform for Rromani could help to optimize future electronic tools for other minority languages.

Keywords :

Rromani, minority languages, dialectology, multi-linguistic, multi-dialectal, e-dictionary, spellchecker