Using Nooj for processing Amazigh language

Fatima Zahra NEJME[1], Siham BOULAKNADEL[2] and Driss ABOUTAJDINE[1]
[1] LRIT, Unité Associée au CNRST (URAC 29), Université Mohamed V-Agdal, Rabat, Maroc
fatimazahra.nejme@gmail.com, aboutaj@fsr.ac.ma
[2] IRCAM, Avenue Allal El Fassi, Madinat Al Irfane, Rabat-Instituts, Maroc
boulaknadel@ircam.ma

Abstract

The Amazigh language in Morocco is considered as a prominent constituent of the Moroccan culture and this by its richness and originality. However it has been long discarded otherwise neglected as a source of enrichment cultural. In the aim of safeguarding its heritage from being threatened of disappearance, it seems opportune to equip this language of necessary means to confront the stakes of access to the domain of New Information and Communication Technologies (ICT). In this context, and in the perspective to build tools and linguistic resources for the automatic processing of Amazigh language, we have undertaken to develop a module for automatic linguistic analysis of the standard Amazigh. To achieve this goal, and given that the linguistic analysis must go through a first step of lexical analysis, which consists in testing membership of each word of the text to the Amazigh lexicon, we started by a formalization of the Amazigh vocabulary namely: noun, verb and particles. According as the noun classification system in Amazigh is computationally interesting and constitutes a major word category which plays a major role in syntactic analysis, we began our study with the formalization of this category, using the finite state technology within the linguistic developmental environment NooJ. The use of this technology was extremely attractive and allows generating and analyzing several thousands of nouns per second.

To achieve this goal, we followed the following steps: (1) construction of electronic dictionary, (2) formalization of the noun category and (3) evaluation.

In the first step, we elaborate our dictionary « EDicAM » (Electronic Dictionary for Amazigh) of simple nouns which contain, actually, 4480 entries presented as singular masculine. This development was based on a set of lexicons: Taifi dictionary (Taifi, 1988), amazigh vocabulary (Ameur et al., 2006b), and media's vocabulary (Ameur et al., 2009a). Each entry into Amazigh dictionary generally presents following details: the lemmas, lexical category, semantic feature and french translation.

In the second step, we have associated each entry in our dictionary with an inflectional class allowing generating all the corresponding inflected forms (feminine, plural and constructed state).

To formalize these classes, we have relied on the works of Boukhris (Boukhris et al., 2008) and those of Oulhaj (Oulhaj, 2000). According to these works and to an heuristic study of the nouns in the Taifi dictionary, we have raised, based on scheme, a set of rules. These rules are structured according:

(1) Gender: 1 rule allowing generating from a masculine noun its feminine correspondent,

(2) Number: 26 rules for the external plural, 29 for the broken plural, 39 rules for mixed one and 1 rule for the plural in ⵙⵏ [id],

(3) State: 9 rules.

After the formalization phase, the finale step was to evaluate our resources. To achieve this goal, we built a test's corpus containing 5524 nouns (in simple and inflected forms) based on Taifi dictionary (Taifi, 1988). The lexical analysis of this corpus shows coverage of about 93% by our resources. The unrecognized nouns mainly due that the noun is inflected with a rule that does not correspond to its exact form, for example the noun ⴰⵎⵎⴰ [aḍḍra] "grain corn", normally the general plural rule for the nouns beginning and ends with ⴰ [a] is to apply an initial vowel alternation accompanied with a suffixation of ⵜⵏ [tn]. However, the noun inflection in the vocabulary is ⴰⵎⵎⴰⵜⵏ [aḍḍratn]. Thus, a set of inflexional problems for gender and state are generated. These types of nouns form the exceptions in Amazigh.