Construction and Analysis of Berber Text Corpus

Khayi Zayd, Olena Orobinska[0000-0001-8396-4136]
National Technical University "Kharkiv Polytechnic Institute",
Kyrpychova str., 2, 61002, Kharkiv, Ukraine

zaydkhayi@gmail.com, orobinska@khpi.edu.ua

Abstract. This work is devoted to constructing a tool to analyze the different aspects of Berber languages. It is based on grammatical parameters of these languages. The text collection containing more than 500 texts that cover long historic period was collected. The corpus is free available and it will useful for further investigations on Tamazigh language. It was transformed into xml-format standardization goal. The corpus counts more than 200 000 of words.
Based on the linguistic rules and statistic methods, original user interface and software prototype were developed by combining the technologies of web-design and object programming in Python.

Keywords: Tamazight language, Corpus Linguistic, Grammar Rules, Statistical Methods.

The grammatical structure of the Berber languages remains poorly understood. There is still no comparative grammar of Berber languages. In order to help overcome this gap, we are constructed the diachronic corpus of the Tamazight languages and we are elaborated the program tool to analyze it. Corpus comprises three major dialects of Moroccan Tamazight: Tarifit, Tachlhit and Tamazight. His size is about 200 000 words and contains more than 520 items of written texts from diverse sources.

When selecting texts, we were guided by the idea that the best way to integrally represent the language is to collect the texts not only in different domains of knowledge but also written in different literary styles. Selected texts were not initially categorized. This work was made in a manual way. Within corpus linguistics, there is currently no commonly accepted approach to the classification of texts. We distinguish 10 categories of texts.

To describe and represent the texts in the corpus we elaborated the XML-structure according to the TEI recommendations.

The interface of the program tool is sown on fig.1.

Using the search function may provide us with type of words we would search for like feminine/masculine nouns and verbs.

Nouns are divided into two parts. The gender in corpus has two forms. The neutral form of word corresponds to masculine while the feminine is indicated by a double t-t

COLINS'2020, Volume II: Workshop. Lviv, Ukraine, April 23-24, 2020, ISSN 2523-4013
http://colins.in.ua, online

affix (the prefix t- and the suffix -t). Ex: Tarbat (girl), Tamtut (woman), Taxamt (tent) and Tislit (bride).

However there are some words whose feminine form contains only the prefix t- and the suffix –a. Ex: Tasa(liver), tawja(family), tarwa (progenitors).

Fig. 1. The interface of MZEGH corpus, English version

Generally, Tamazight masculine words have such prefixes that distinguish them from other words. For instance, 'a', 'u', 'i'. Ex: Asklu (tree) udi (cheese) ighef (head). As differentiated from the rule for feminine nouns, the rule for the defining the masculine nouns has a fair bit of exeptions.

Verbs in corpus are for the first person singular and plural that has suffixes 'agh','ex', 'egh'. Ex:'ghrex' (I study), 'fegh' (I go out), 'nadagh' (I call).

The program tool permits to obtain such characteristics of this corpus:

- list of all tokens;
- list of unique words;
- lexical diversity;
- realize different grammatical requests/

Actually we are working on the method which enables to automatically group the words in grammatical classes like NOUN, VERB, ADJECTIV using n-gram method.

References

1. Sinclair, J.. (2004). Trust the text: Language, corpus and discourse. 1-212. 10.4324/9780203594070.
2. Brenier-Estrine, C., Institut de recherches et d'études sur le monde arabe et musulman. (1994). Bibliographie berbère annotée: 1992-1993. Aix-en-Provence: Institut de recherches et d'études sur le monde arabe et musulman.