World Academy of Science, Engineering and Technology
International Journal of Information and Communication Engineering
Vol:18, No:03, 2024

Construction and Analysis of Tamazight (Berber) Text Corpus

Authors : Zayd Khayi

Abstract : This paper deals with the construction and analysis of the Tamazight text corpus. The grammatical structure of the Tamazight remains poorly understood, and a lack of comparative grammar leads to linguistic issues. In order to fill this gap, even though it is small, by constructed the diachronic corpus of the Tamazight language, and elaborated the program tool. In addition, this work is devoted to constructing that tool to analyze the different aspects of the Tamazight, with its different dialects used in the north of Africa, specifically in Morocco. It also focused on three Moroccan dialects: Tamazight, Tarifiyt, and Tachlhit. The Latin version was good choice because of the many sources it has. The corpus is based on the grammatical parameters and features of that language. The text collection contains more than 500 texts that cover a long historical period. It is free, and it will be useful for further investigations. The texts were transformed into an XML-format standardization goal. The corpus counts more than 200,000 words. Based on the linguistic rules and statistical methods, the original user interface and software prototype were developed by combining the technologies of web design and Python. The corpus presents more details and features about how this corpus provides users with the ability to distinguish easily between feminine/masculine nouns and verbs. The interface used has three languages: TMZ, FR, and EN. Selected texts were not initially categorized. This work was done in a manual way. Within corpus linguistics, there is currently no commonly accepted approach to the classification of texts. Texts are distinguished into ten categories. To describe and represent the texts in the corpus, we elaborated the XML structure according to the TEI recommendations. Using the search function may provide us with the types of words we would search for, like feminine/masculine nouns and verbs. Nouns are divided into two parts. The gender in the corpus has two forms. The neutral form of the word corresponds to masculine, while feminine is indicated by a double t-t 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), and tarwa (progenitors). Generally, Tamazight masculine words have prefixes that distinguish them from other words. For instance, 'a', 'u', 'i', ex: Asklu (tree), udi (cheese), ighef (head). Verbs in the corpus are for the first person singular and plural that have suffixes 'agh','ex', 'egh', ex: 'ghrex' (I study), 'fegh' (I go out), 'nadagh' (I call). The program tool permits the following characteristics of this corpus: list of all tokens; list of unique words; lexical diversity; realize different grammatical requests. To conclude, this corpus has only focused on a small group of parts of speech in Tamazight language verbs, nouns. Work is still on the adjectives, prounouns, adverbs and others.

Keywords : Tamazight (Berber) language, corpus linguistic, grammar rules, statistical methods
Conference Title : ICCALS 2024 : International Conference on Communication and Linguistics Studies
Conference Location : Barcelona, Spain
Conference Dates : March 04-05, 2024