Towards Electronic Lexicography for the Kurdish Language

Sina Ahmadi, Hossein Hassani, John P. McCrae
sina.ahmadi@insight-centre.org

Context
A machine-readable dictionary (MRD) not only provides lexicographic information in an electronic form, but is also a database which can be queried and therefore integrated in natural language processing tools. As the body of the research in Kurdish language processing is still scant, we believe that such resources will pave the way for further developments in the field. We also believe that lexical resources will enable researchers to address more NLP tasks which may require lexicographic resources such as word sense disambiguation and semantic parsing and, enhance the quality of the existing NLP applications.

The Kurdish language
Kurdish is
▶ an Indo-European language spoken by about 30 million speakers
▶ spoken in several dialects, such as Kurmanji, Sorani, Hawrami and Kirmashani
▶ written using different scripts, such as Persian-Arabic, Latin and Cyrillic
▶ less-resourced, i.e. general-purpose grammars and raw internet-based corpora are the main existing resources

Objectives
▶ We provide a review of the current state of Kurdish lexicography, both traditional and electronic, including an analysis of the properties of the existing Kurdish dictionaries, such as type of dictionary (monolingual, bilingual, multilingual), script of the Kurdish text (Persian-Arabic, Latin or Cyrillic), description of the content and size of dictionaries.
▶ We present three machine-readable dictionaries based on the OntoLex-Lemon model for Kurmanji, Sorani and Hawrami dialects.

There are reportedly more than 71 printed dictionaries and terminological resources available for Kurdish. An analysis of the resources to which we could have access, i.e. 60 resources, is provided in the following Figure:

Methodology
In order to create our dictionaries, we followed the pipeline illustrated in the following Figure:

In addition to OntoLex-Lemon core, we used the following modules: lime, synsem, lexinfo, vartrans and lexicog.

Evaluation
We selected three dictionaries [1,2,3] for our experiment following three selection criteria:
▶ the number of entries to be manageable in a research project
▶ the availability of the resource
▶ the copyright situation of the resource.

Figure: An example entry from our Kurmanji-English dictionary. The original printed entry versus the equivalent in RDF Turtle based on the OntoLex-Lemon model

Our resources are available at /KurdishBLARK/KurdishLex.

References

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.