Towards Electronic Lexicography for the Kurdish Language

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

Insight
Centre for Data Analytics

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:

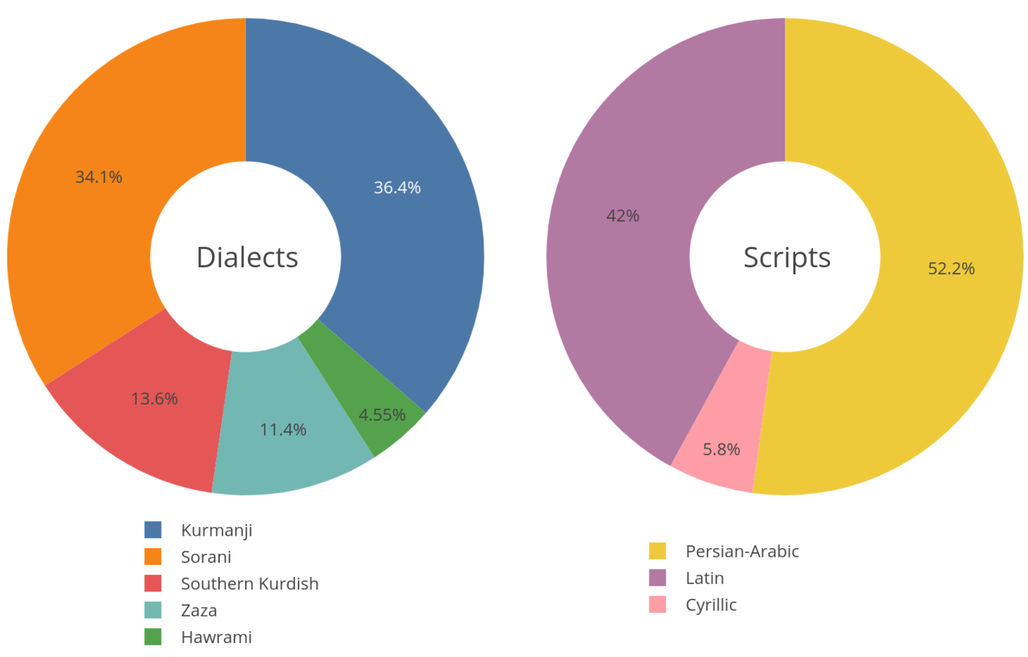
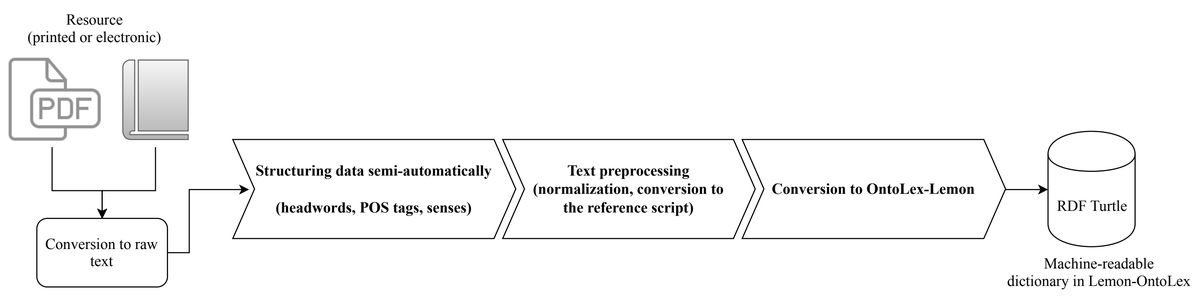


Figure: Distribution of Kurdish lexicographic resources across dialects and based on scripts

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**.

bend *f* bond; **li ~a** for the sake of, chained to, waiting for: *divê em êdî li benda sibehê ranewestin* we shouldn't stand around waiting for tomorrow; **~ kirin** *v.t.* to fetter, arrest; **man di ~a** to wait for

```
:lexicon a lime:Lexicon;
    lime:language <www.lexvo.org/page/iso639-3/kmr> ;
    lime:entry :lex_bend .

:lex_bend a ontolex:LexicalEntry, ontolex:Word ;
    ontolex:canonicalForm :form_bend ;
    rdfs:label "bend"@kmr-latn .

:form_bend a ontolex:Form ;
    dct:language <www.lexvo.org/page/iso639-3/kmr> ;
    ontolex:writtenRep "bend"@kmr-latn ;
    lexinfo:partOfSpeech lexinfo:noun ;
    lexinfo:gender lexinfo:feminine ;
    lexinfo:number lexinfo:singular ;
    ontolex:sense :bend_n_sense .

:en_bond a ontolex:LexicalEntry ;
    dct:language <http://lexvo.org/id/iso639-1/en> ;
    ontolex:sense :en_bond_sense .

:trans a vartrans:Translation ;
   vartrans:source :bend_n_sense ;
   vartrans:target :en_bond_sense .

:bend_n_sense a lexicog:UsageExample ;
   rdf:value "divê em êdî li benda sibehê
       ranewestin."@kmr-latn ;
   rdf:value "we shouldn't stand around waiting for
       tomorrow."@en .
```

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

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.

Resource	Number of entries		Attributes				Polysemy degree
	Word	MWE	Gender & POS	Etymology	# idioms	Examples	
Kurmanji	4172	122	3420 (76.64%)	213 (4.96%)	340	265 (6.35%)	1.03%
Sorani	5683	160	5348 (91.37%)	111 (1.89%)	82	543 (9.55%)	1.06%
Hawrami	1184	165	1184 (87.76%)	242 (17.93%)	123	10 (0.008%)	1.01%

Table: Lexicographic resources statistics

Our resources are available at /KurdishBLARK/KurdishLex.

References

1. Thackston, W.M. (2006a). Kurmanji Kurdish–A Reference Grammar with Selected Readings. Harvard University.
2. Thackston, W.M. (2006b). Sorani Kurdish—A Reference Grammar with Selected Readings. Harvard University.
3. MacKenzie, D.N. (1966). The dialect of Awroman (Hawraman-i Luhon) Grammatical sketch, texts, and vocabulary. Copenhagen: Munksgaard.

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

elexis european lexicographic infrastructure
NUI Galway OÉ Gaillimh
Science Foundation Ireland SFI For what's next
University of KURDISTAN Hewlêr