KLPT – Kurdish Language Processing Toolkit

Sina Ahmadi

Insight Centre for Data Analytics, National University of Ireland Galway, Ireland

Objectives

- Describe the **Kurdish language** as a less-resource language
- Carry out **a survey** on what has been done in Kurdish language processing
- Present **a toolkit in Python** for Kurdish language processing
- Address **basic language processing tasks** for Kurdish

Introduction

Despite the recent advances in applying **language-independent approaches** to various natural language processing tasks thanks to artificial intelligence, some **language-specific tools** are still essential to process a language in **a viable manner**.

Although there is a plethora of performant tools and specific frameworks for NLP, such as NLTK, Stanza and spaCy, the progress with respect to less-resourced languages is often hindered by not only the **lack of basic tools and resources** but also the **accessibility of the previous studies** under an open-source license.

The Kurdish language

Figure 1:Kurdish settlements in Southwest Asia (Encyclopædia Britannica)

- Kurdish is a **less-resourced Indo-European language** spoken by 20-30 million speakers in the Kurdish regions of Iran, Iraq, Turkey and Syria [1]

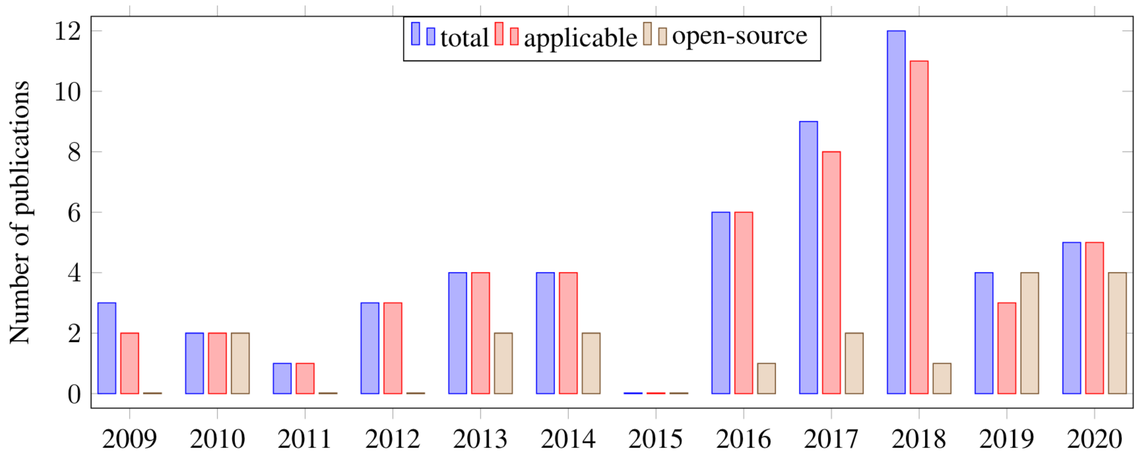


Figure 2:Number of scientific publications in KLP

- Northern Kurdish (Kurmanji), Central Kurdish (Sorani), Southern Kurdish and Laki are the dialects of Kurdish
- An **Arabic-based alphabet** is widely used for Sorani and Southern Kurdish while a **Latin-based** is used for Kurmanji
- There is no consensus regarding what is meant by a **standard writing system** or **orthography**
- Kurdish language processing (KLP) is still in the early stages of development

Current state of KLP

We reviewed the scientific publications that directly address an issue in those fields. We analyze previous publications from the following two perspectives:

- **Applicability**: Does the paper propose an approach or methodology that can be applied to solve the same problem in the other dialects of Kurdish? the majority
- **Open-source**: Does the paper provide the discussed resource or tool under an open-source license? only 18 out of 53

Kurdish is still a less-resourced language

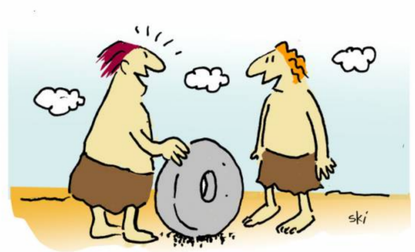
KLPT modules

KLPT is implemented in **Python** and is composed of **four core modules** with specific tasks without using any external NLP library:

- `Preprocess`: Handle diversities in scripts and orthographies in a formalized way
- `Transliterate`: transliterating the Arabic based and Latin based scripts of Kurdish
- `Stem`: spelling error detection and correction, stemming, lemmatization and morphological analysis
- `Tokenize`: tokenize words, sentences and multi-word expressions

Lessons learned

- **Don't reinvent the wheel**: Release your project under an open source license
- Create **community-driven** initiatives
- **Raise awareness** by promoting good practices in content creation on the Web



Demonstration

```
>>> from klpt.preprocess import Preprocess
>>> from klpt.transliterator import Transliterate
>>> from klpt.tokenize import Tokenize
>>> from klpt.stem import Stem

# Preprocess module
>>> preprocessor = Preprocess("Sorani", "Arabic",
numeral="Latin")
>>> preprocessor.normalize("لە ساڵەکانی ١٩٥٠دا")
لە ساڵەکانی 1950دا
>>> preprocessor.standardize("راستە لەو وڵاتەدا")
ڕاستە لەو وڵاتەدا

# Transliterate module
>>> transliterator = Transliterate("Kurmanji", "Latin",
target_script="Arabic")
>>> transliterator.transliterate("rojhilata navîn")
'رۆژهلاتا نافین'

# Stem module
>>> stemmer = Stem("Sorani", "Arabic")
>>> stemmer.check_spelling("سوتاندبووت")
False
>>> stemmer.correct_spelling("سوتاندبووت")
('سووتاندبووت', 'سووتاندەت', 'سووتاندن', 'سووتاند')
>>> stemmer.stem("سووتاندبووت")
('سووت',)
>>> stemmer.analyze("دیتبامن")
{'pos': 'verb', 'is': 'past_intransitive', 'stem':
'دی', 'verb_stem': 'دیت', 'terminal_suffix': 'بامن'}

# Tokenize module
>>> tokenizer = Tokenize("Kurmanji", "Latin")
>>> tokenizer.word_tokenize("endamên encûmena wezîrên")
['_endam_ên', '_encûmen_a', '_wezîr_ên']
```

References

[1] Sina Ahmadi, Hossein Hassani, and John P. McCrae. Towards Electronic Lexicography for the Kurdish Language.
In *Proceedings of the eLex 2019 conference*, pages 881–906, Sintra, Portugal, 1–3 October 2019. Brno: Lexical Computing CZ, s.r.o.

Use the tool

- Toolkit:
  `https://github.com/sinaahmadi/klpt`
- Documentation:
  `https://sinaahmadi.github.io/klpt/`