Technology for Minoritized Language Communities

An Overview of Language and Speech Technology for Kurdish

Sina Ahmadi
George Mason University
https://sinaahmadi.github.io

University of Toronto
May 25, 2023

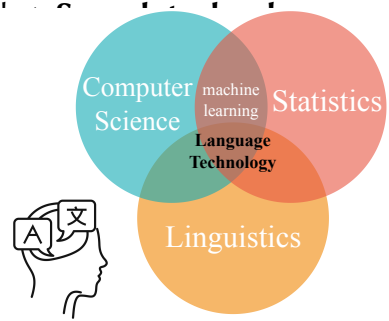
Table of Contents

Language and Speech Technology: What Is It?

- **Theoretical Linguistics**: Nature of language
  Phonetics, phonology, morphology, syntax etc
- **Sociolinguistics**: Languages and social factors
  "... one giant leap for *mankind*." → ~~*womankind*~~?
- **Psycholingusitics**: Languages and psychology
  What actually is a "thought"?
- **Applied linguistics**: Real-life applications
  Teaching and using language
- **Historical linguistics**: evolution of language
  'Nice' (adjective) used to mean 'ignorant'!
- **Computational linguistics**: languages and computers ⇒ **our topic today**

LST: “understanding” language computationally

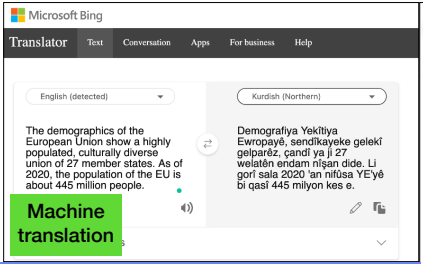
- **Computational linguistics (CL)**: the study of languages using computational techniques. It is about *linguistics*.
- **Natural language processing (NLP)**: the creation of tools, algorithms and resources to solve tasks related language processing. It is about *engineering*.
- **CL** and **NLP** are often conflated and used interchangeably.
- Language as text ⇒ **Language technology**
- Language a

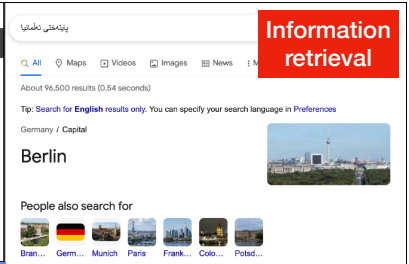


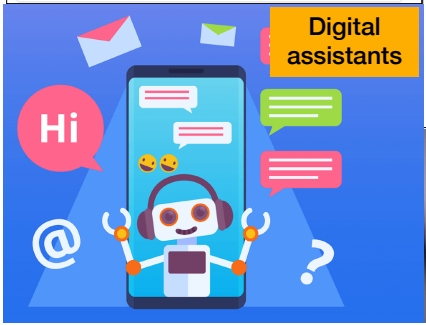
Computational Linguistics (CL)

Natural Language Processing (NLP)

Language and Speech Technology: a few applications

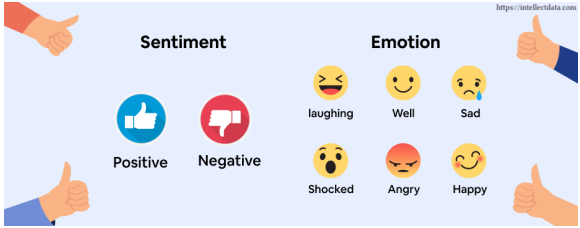






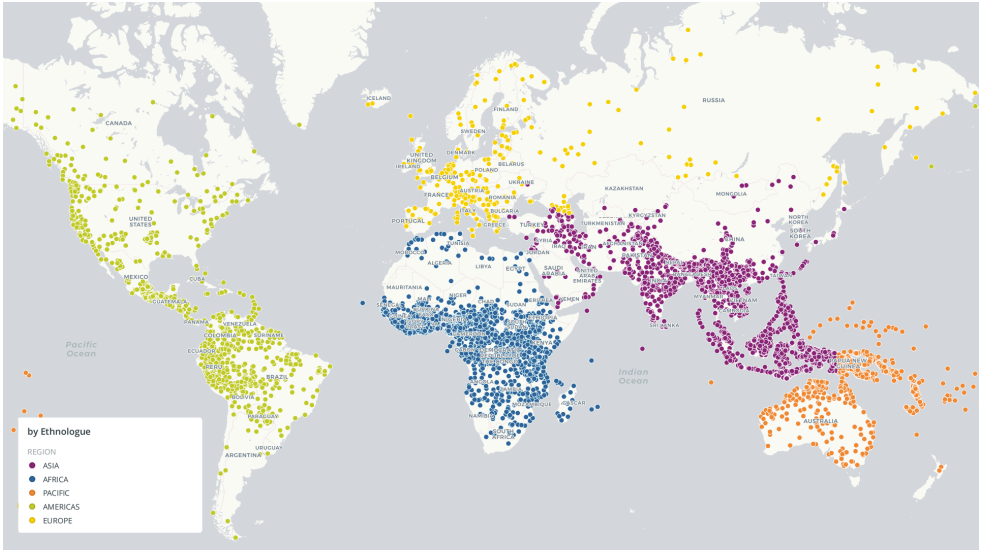
Language and Speech Technology: a few tasks

- Machine translation
- Word-sense disambiguation
- Spelling error correction
- Question answering
- Syntactic parsing
- Text summarization
- Sentiment & emotion analysis
- And many more...

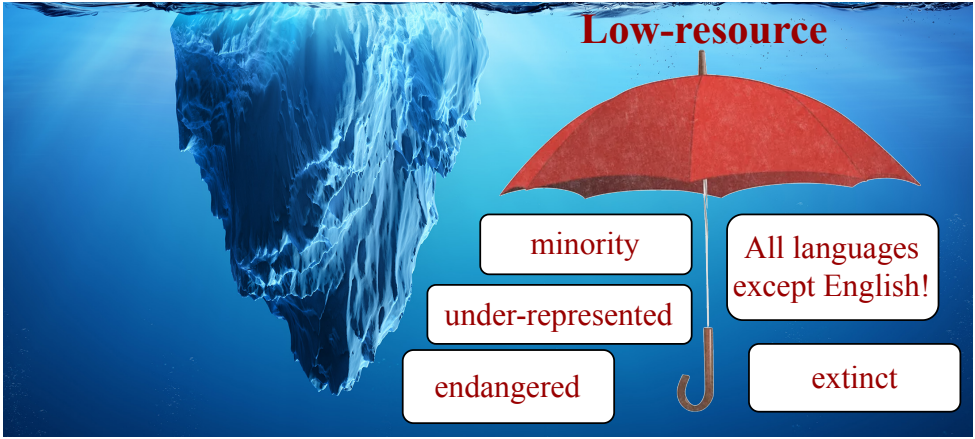


Language and Speech Technology: Linguistic Disparity

- More than 7,000 “languages” are spoken today (Ethnologue, 2023).



Language and Speech Technology: Linguistic Disparity

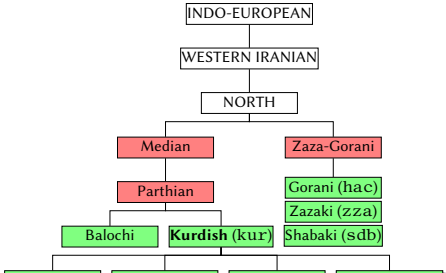


- **99%** of languages around the globe are low-resourced, including Kurdish!

Table of Contents

Kurdish Language

- an Indo-European language
- spoken by 20-30 million speakers
- spoken in many dialects and subdialects (*dialects* or *languages*?)
- has a longer oral tradition than a written one $\Rightarrow$ *lack of data*
- written in many scripts: the Latin-based and Arabic-based ones still widely in use



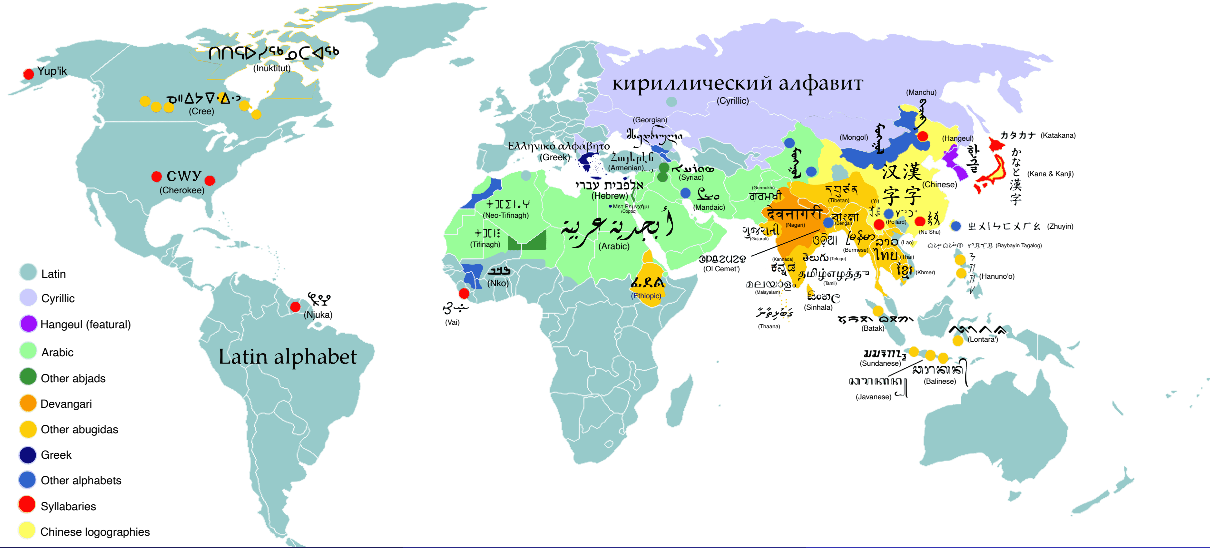
Source: https://www.britannica.com/topic/Kurd

Kurdish Language: Issues

- Too many scripts scatter readers and creates further challenges in NLP
- Written in various orthographies following different conventions
- Even though, still written *unconventionally*
  - *di sala 2020'an* | *2020-an* | *2020an de* "in the year 2020"
  - *hêv**i**ya*, *hêv**î**ya* or *hêvî ya* "hope of"?
  - ٠١٢٣٤٥٦٧٨٩, ٠١٢٣٤٥٦٧٨٩ or 0123456789?
- Kurdish orthographies are phonemic, but not always:
  - double-usage characters: ی for î/y and و for u/w
  - variations like ł, ll or ɬ for [ɬ]
  - vowel *i* missing in the Arabic-based

Kurdish Language: Unconventional Writing

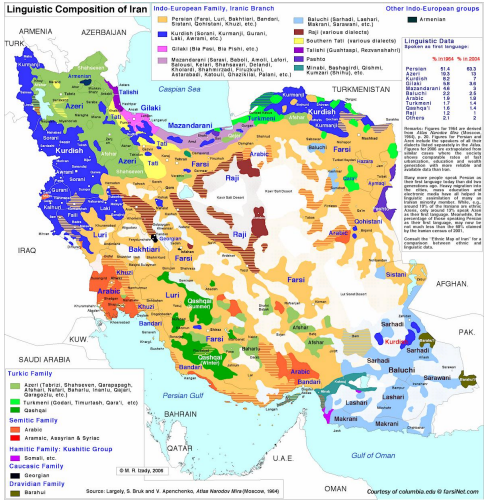
- Almost 300 writing systems exist (and many adopted ones)
- Less than 4,000 languages have a written form



Kurdish Language: Unconventional Writing

Most countries are X-lingual, but not all officially!

- Turkey:
  → Turkish
- Syria:
  → Arabic
- Iraq:
  → Arabic and Kurdish
- Iran:
  → Persian



Kurdish Language: Unconventional Writing

Unconventional writing

Unconventional writing refers to the usage of the **script of another language**, presumably that of a dominant language.

- Unsystematic writing
- Not necessarily complying with orthography
- Impact of donor language on code switching
- No specific rule for mapping graphemes-phonemes

A few examples:

- *ana raye7 el gam3a el sa3a 3 el 3asr.* (Ararbic in Latin script, aka Arabizi)
- *Tora ti na sou po* (Greek in Latin, aka Greeklish)
- *mer6 pr tn mess pr mn anif* (French SMS language)

Kurdish Language: Unconventional Writing

Language	Unconventional script	Unconventional writing	Conventional writing
Gilaki	Persian	یته زون نم هیسه گه گیلکن اون جی گب زنن	یته زوؤن ٚ نؤم هیسه گه گیلکؤن اؤن ٚ جي گب زنن
Kashmiri	Urdu	برور چھ اکھ وراسے جانور۔	برٛور چھُ اَکھ وُرأسۍ جانوَر۔
Kurmanji	Arabic	قایمقام اامدي بةرثوا پارزكار دهوك دا	قایمقامێ ئامێدیێ بەرسڤا پارێزگارێ دهۆکێ دا
Sorani	Arabic	هةر لة يةكةم شانؤوة دياره فهديان دةويت	هەر لە یەکەم شانۆوە دیارە فەهەدیان دەوێت
Sindhi	Urdu	مديني ڏانهن هجرت وقت فقط هيءَ گهرواري سائن گڏ هئي	مدیني ڏانهن هجرت وقت فقط هيءَ گهرواري سائن گڏ هئي

Kurdish Language: Complex Morphology

- Kurdish has a synthetic morphology: $>$ 2000 noun forms from a stem!
- More synthetic than Old English and Yakut and less than Sanskrit
- Complex morphotactics due to split-ergativity

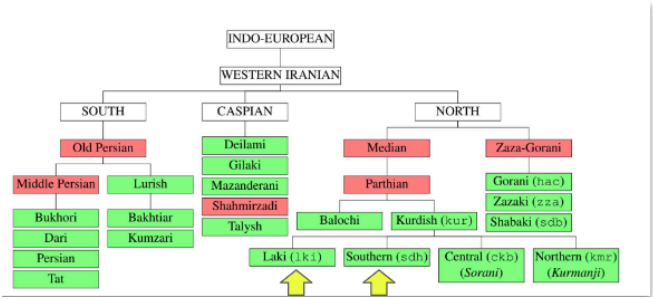
0					girt						past stem of GIRTIN (to take, to get)
1					girt	im					I got.
2					girt	im	in				I got them.
3					girt	im	in	e			I got them to/with.
4					girt	im	in	e	ewe		I got them to/with again.
5					girt	îş	im	in	e	ewe	I got them also to/with again.
6			ne	îş	im	girt	in	e	ewe		I did not get them also to/with again.
7		ne	îş	im	de	girt	in	e	ewe		I was not getting them also to/with again.
8	da	îş	im	ne	de	girt	in	e	ewe		I was not taking down them also to/with again.

The placement of the endoclitic *=îş* (in green boxes) and agent marker *=im* (in blue boxes) with respect to the base and each other in a verb form. Note that Sorani Kurdish is a null-subject language.

Kurdish Language: Underrepresented Varieties

Not too many resources available for Southern Kurdish and Laki:

- Spoken by >2M in western Iran, border regions of Iraq, and its capital, Baghdad
- The classification of Laki is still debated
- Faced various discriminatory language policies leading to pernicious sociolinguistic effects
- Lack of children proficiency in Southern Kurdish and limited usage of the language in writing
- Few available digital resources available [Ahmadi et al., 2019]



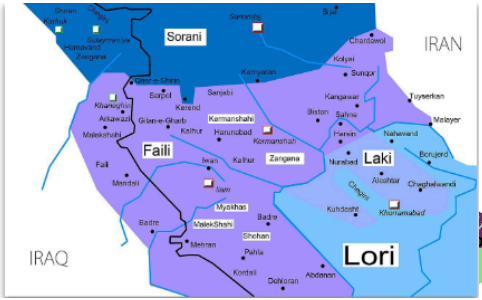


Table of Contents

Kurdish Language Processing: Scientific Contributions

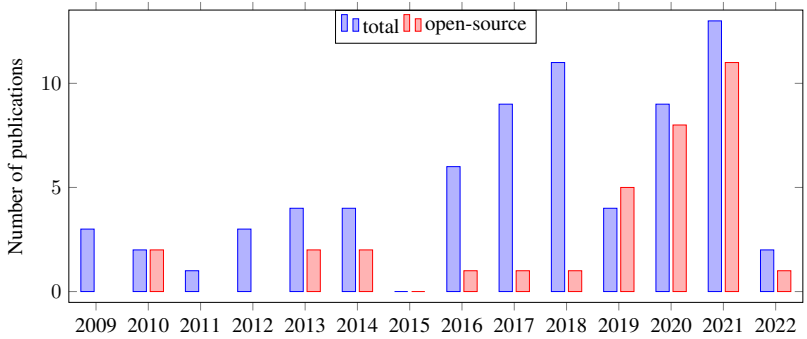
Less than 100 publications address a task in Kurdish language technology.

- the earliest works in the field of KLP date back to 2009
  except a work in 1995 [Baban and Husein, 1995] – Why such an interval?!)
- thus far, a total number of **87** publications are published in a field directly related to KLP, including non-peer-reviewed ones
- all varieties have not equally received attention

Open-source

Does the paper provide the discussed resource or tool under an open-source license?

Kurdish Language Processing: Scientific Contributions



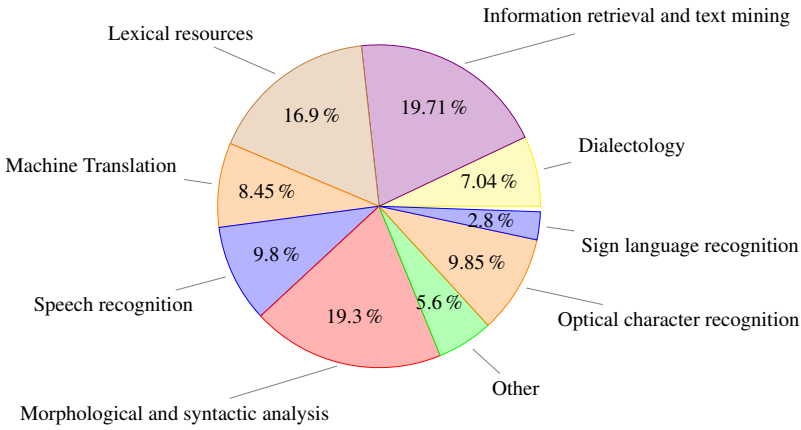
Number of scientific publications directly related to KLP per year and field

- Roughly a third provide their resources or tools under an open-source license
- Central Kurdish makes up a predominant proportion of almost 90% of publications
- Only one publication addresses the processing of Southern Kurdish, Laki or Zazaki

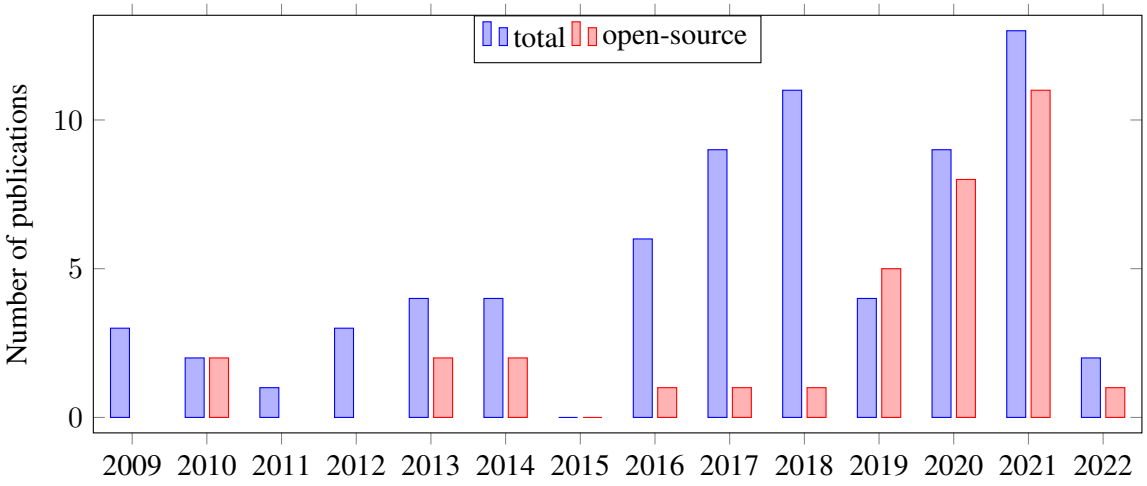
Kurdish Language Processing: Scientific Contributions

A range of topics have been addressed:

- Creation of lexical resources and corpora [Ahmadi et al., 2023]
- Morphological and syntactic analysis [Ahmadi, 2020c]
- Sentiment analysis [Hameed et al., 2023]
- Language identification [Ahmadi et al., 2023]
  - And many more ⇒ **https://github.com/sinaahmadi/awesome-kurdish**



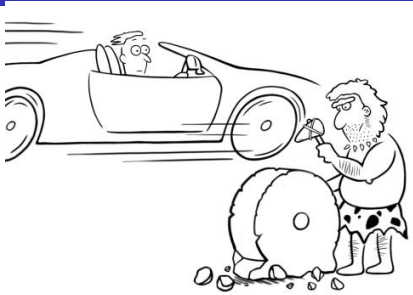
Kurdish Language Processing: What is wrong?



Number of scientific publications directly related to KLP per year and field

Kurdish Language Processing: What is wrong?

- Many projects overlap significantly, yet none of them provide a solution under any open-source license
  - Stemming is addressed at least *six* times [Jaff, 2014, Salavati and Ahmadi, 2018, Mustafa and Rashid, 2018, Saeed et al., 2018, Hawezi et al., 2019, Ahmadi, 2020e]



- Some are hardly integrable or inter-operable
  - A large-scale morphological lexicon and a part-of-speech tagger for Kurdish within the Alexina framework [Walther and Sagot, 2010]
- Released in an unorganized manner for individual tasks
  - Example: a transliteration tool for Kurdish [Ahmadi, 2019a]
- Under-represented variants
- A lack of involvement of the Kurdish linguists
- No funded project
- **Kurdish *is* still a less-resourced language**

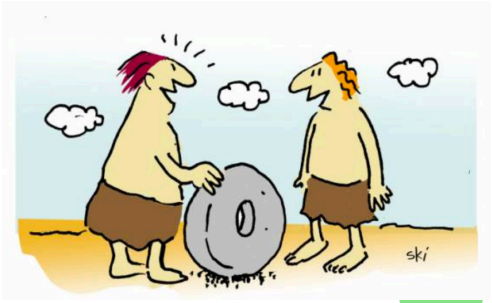
Kurdish Language Processing: a few projects

- Wîkîferheng: https://ku.wiktionary.org
- Wîkîpediyaya kurdî: https://ku.wikipedia.org
- VejînLex: https://lex.vejin.net/en
- Importance of Kurdish for the Tech Giants
  - Google: https://translate.google.com
  - Microsoft: https://www.bing.com/translator
  - Meta: https://ai.facebook.com/research/no-language-left-behind
- Kurdish Computational Linguistics Course: https://sinaahmadi.github.io/KurdishCL

KLP: Kurdish Language Processing Toolkit

- a basic but extendable language processing toolkit
- an effort to standardize Kurdish language with all its dialects and scripts
- implemented in Python
- inspired by the functionality of relevant NLP toolkits, e.g. NLTK and spaCy
- no external NLP library is used in this toolkit
- composed of core modules for Sorani and Kurmanji for the following tasks:
  - text preprocessing
  - stemming
  - lemmatization
  - spelling error detection and correction
  - transliteration
  - morphological analyzer and generator
  - tokenization
- it is open-source!

  → **https://github.com/sinaahmadi/klpt**



KLP: Which tasks to be addressed next?

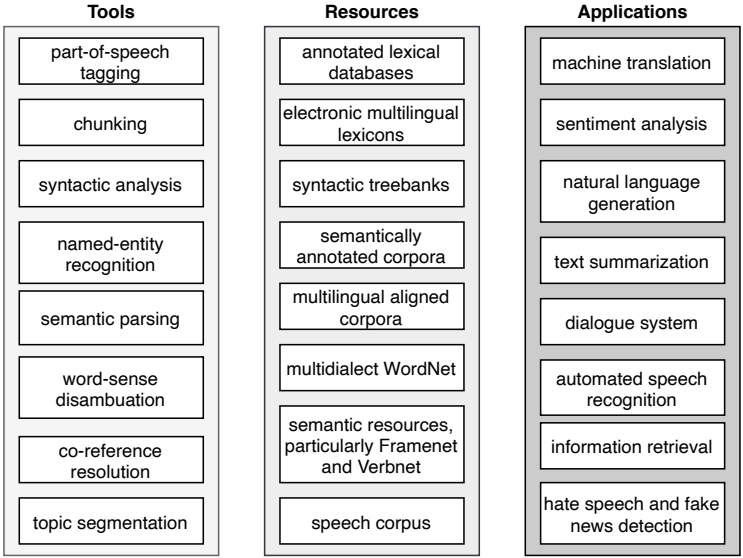


Table of Contents

Conclusion

- **Lessons learned**:
  - **release your project under an open source license** → essential to ensure gradual but efficient progress in resource and technology development for a less-resourced language
  - **community-driven initiatives** → bring together users, developers, researchers, language activists and policy makers
  - **raise awareness** by promoting good practices in content creation on the Web, particularly collaboratively-curated resources such as Wiktionary[1] and Wikipedia[2]
  - every single user is a contributor too
  - **time to reconcile linguistics with computational methods for Kurdish**
- **Future directions**:
  - promote and extend technology for KLP
  - create a community of developers and linguists for KLP
  - train future professors and researchers in the field ⇒ needs $$$

[1] https://en.wiktionary.org

[2] https://www.wikipedia.org/

And, the takeaway point is ...

"An endangered language will progress if its speakers can make use of electronic technology."
– David Crystal (Language death, p.13)

Recommended readings

- Linguistics
  - Routledge book series on linguistics (link)
- Computational Linguistics
  - Speech and Language Processing (link)
  - The Handbook of CL & NLP (link)
- Kurdish Linguistics
  - Sorani & Kurmanji reference grammars (W. M. Thackston) (link)
  - Kurdish dialect studies (Mackenzie, D. N.) (link)
- Programming in Python
  - https://www.learnpython.org
  - Natural Language Processing with Python (link)

Thanks!

Spas!

https://github.com/sinaahmadi/klpt

KLPT

References

Sardar Jaf, Allan Ramsay (2014)

Stemmer and a POS tagger for Sorani Kurdish.

6th International Conference on Corpus Linguistics - Spain.

Shahin Salavati and Sina Ahmadi (2018)

Building a Lemmatizer and a Spell-checker for Sorani Kurdish.

arXiv preprint arXiv:1809.10763.

Mustafa, Arazo M., and Tarik A. Rashid. (2018)

Kurdish stemmer pre-processing steps for improving information retrieval

Journal of Information Science, 44.1: 15-27.

Saeed, A. M., Rashid, T. A., Mustafa, A. M., Agha, R. A. A. R., Shamsaldin, A. S., & Al-Salihi, N. K. (2018)

An evaluation of Reber stemmer with longest match stemmer technique in Kurdish Sorani text classification

Iran Journal of Computer Science, 1(2), 99-107.

Hawezi, R. S., Azeez, M. Y., & Qadir, A. A. (2019)

Spell checking algorithm for agglutinative languages Central Kurdish as an example

International Engineering Conference (IEC)(pp. 142-146). IEEE.

Sina Ahmadi (2019)

A Rule-based Kurdish Text Transliteration System

Asian and Low-Resource Language Information Processing (TALLIP) 18(2):18:1–18:8.

Sina Ahmadi (2020)

A Tokenization System for the Kurdish Language

Proceedings of the Seventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2020).

Sina Ahmadi (2020)

A Formal Description of Sorani Kurdish Morphology

`https://arxiv.org/abs/2109.03942`.

Sina Ahmadi (2020)

Building a Corpus for the Zaza–Gorani Language Family

Proceedings of the Seventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2020).

Sina Ahmadi (2020)

Hunspell for Sorani Kurdish Spell checking and Morphological Analysis.

`https://arxiv.org/abs/2109.06374`.

Walther, G., & Sagot, B. (2010)

Developing a large-scale lexicon for a less-resourced language: General methodology and preliminary experiments on Sorani Kurdish.

7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC 2010 Workshop).

Baban, ST and Husein, S (1995)

Programmable Grammar of the Kurdish Language

ILLC Research Report and Technical Notes.

Ahmadi, Sina and Hassani, Hossein and McCrae, John P (2019)

Towards electronic lexicography for the Kurdish language

Proceedings of the sixth biennial conference on electronic lexicography (eLex).

Hameed, Razhan and Ahmadi, Sina and Daneshfar, Fatemeh (2023)

Transfer Learning for Low-Resource Sentiment Analysis

arXiv preprint arXiv:2304.04703.

Sina Ahmadi, Milind Agarwal and Antonios Anastasopoulos (2023)

PALI: A Language Identification Benchmark for Perso-Arabic Scripts

arXiv preprint arXiv:2304.01322.

Sina Ahmadi, Zahra Azin, Sara Belelli and Antonios Anastasopoulos (2023)

Approaches to Corpus Creation for Low-Resource Language Technology: the Case of Southern Kurdish and Laki

arXiv preprint arXiv:2304.01319.