

# 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



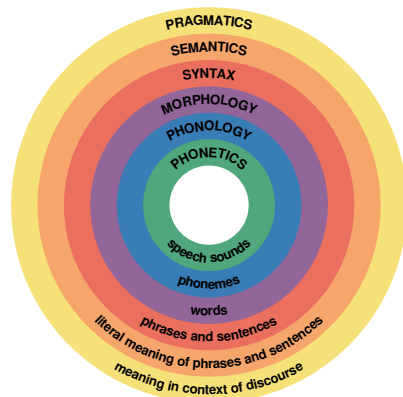
# Table of Contents

- 1 Language and Speech Technology
- 2 Kurdish Language
- 3 Kurdish Language Processing (KLP)
- 4 Conclusion



# Language and Speech Technology: What Is It?

- **Theoretical Linguistics:** Nature of language  
Phonetics, phonology, morphology, syntax etc



# 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*?



# 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*?
- **Psycholinguistics:** Languages and psychology  
What actually is a “thought”?



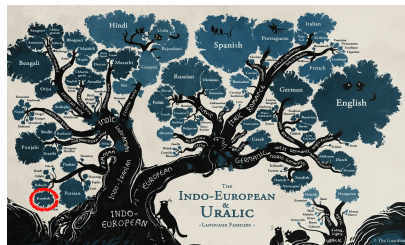
# Language and Speech Technology: What Is It?

- **Theoretical Linguistics:** Nature of language  
Phonetics, phonology, morphology, syntax etc
- **Sociolinguistics:** Languages and social factors  
“... one giant leap for *mankind*.” → *womankind*?
- **Psycholinguistics:** Languages and psychology  
What actually is a “thought”?
- **Applied linguistics:** Real-life applications  
Teaching and using language



# 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*?
- **Psycholinguistics:** 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’!



# Language and Speech Technology: What Is It?

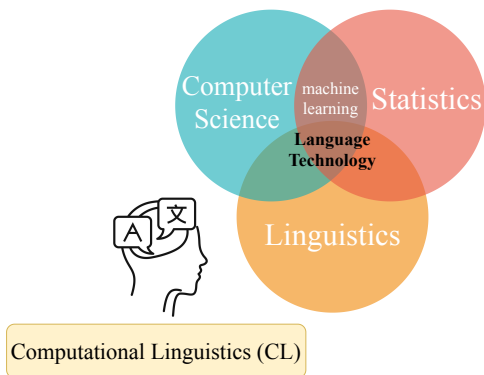
- **Theoretical Linguistics:** Nature of language  
Phonetics, phonology, morphology, syntax etc
- **Sociolinguistics:** Languages and social factors  
“... one giant leap for *mankind*.” → *womankind*?
- **Psycholinguistics:** Languages and psychology  
What actually is a “thought”?
- **Applied linguistics:** 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**





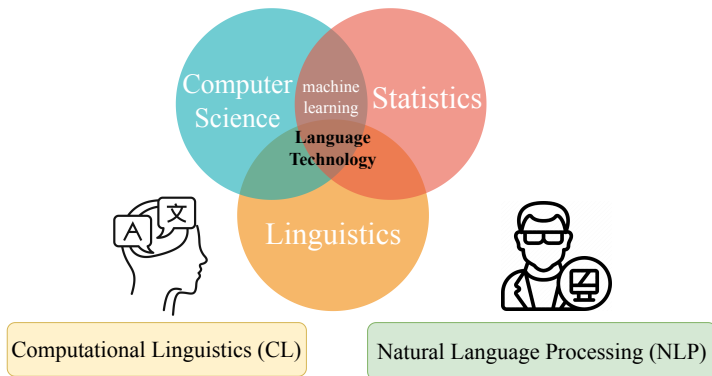
# LST: “understanding” language computationally

- **Computational linguistics (CL):** the study of languages using computational techniques. It is about *linguistics*.



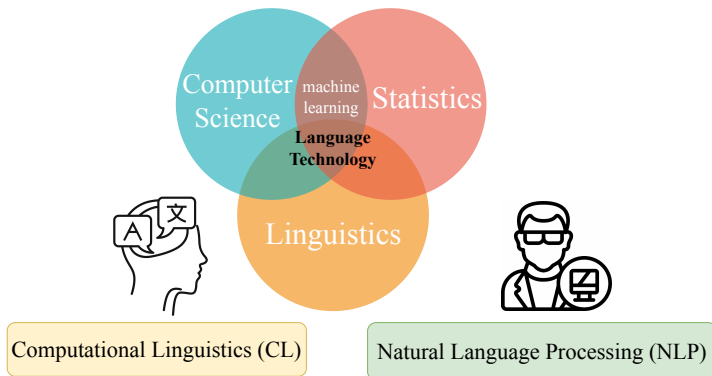
# LST: “understanding” language computationally

- **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*.



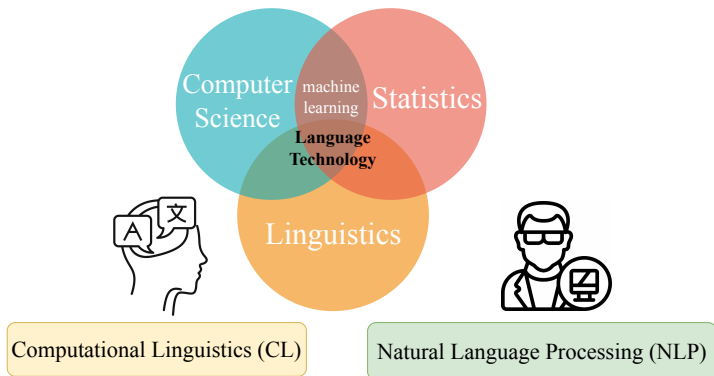
# LST: “understanding” language computationally

- **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.



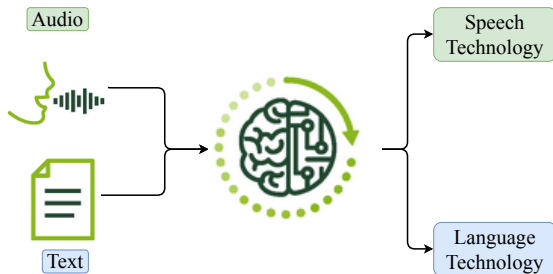
# LST: “understanding” language computationally

- **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  $\Rightarrow$  **Language technology**

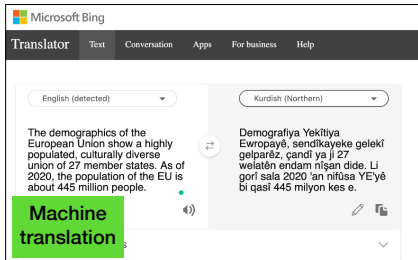


# LST: “understanding” language computationally

- **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  $\Rightarrow$  **Language technology**
- Language as sound  $\Rightarrow$  **Speech technology**



# Language and Speech Technology: a few applications



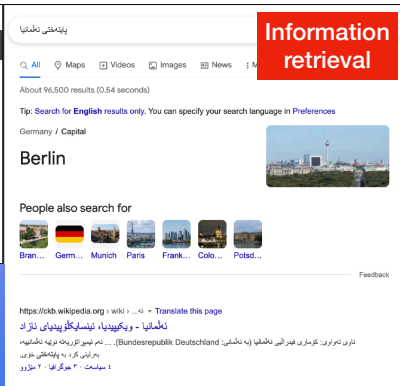
Microsoft Bing  
Translator Text Conversation Apps For business Help

English (detected) Kurdish (Northern)

The demographics of the European Union show a highly populated, culturally diverse union of 27 member states. As of 2020, the population of the EU is about 445 million people.

Demografiya Yekîtiya Ewropayê, sendikayê gelekî gelparêz, çandî ya ji 27 welatên endam nîşan dide. Li gorî sala 2020'an nîfûsa YE'yê bi qasî 445 milyon kes e.

**Machine translation**



پاتەختی ئەڵمانیا

**Information retrieval**

All Maps Videos Images News

About 96,500 results (0.54 seconds)

Tip: Search for **English** results only. You can specify your search language in Preferences

Germany / Capital

## Berlin

People also search for

Bran... Germ... Munich Paris Frank... Colo... Potsd...

Feedback

<https://ckb.wikipedia.org/wiki/ئەڵمانیا> Translate this page

ئەڵمانیا - ویکیپیدیا، نۆڤسەرکۆڵێنێکی زۆرێک لە زۆرێک  
ئەڵمانیا (بە ئەڵمانی: Bundesrepublik Deutschland) ... نام نۆڤسەرکۆڵێنێکی زۆرێک لە زۆرێک،  
بەرامبەرێکی بە پاتەختی خۆی.  
ئەڵمانیا - 3 سۆڤسەت - 2 مۆڤرگیا - 2 مۆڤر



**Digital assistants**

Hi

@

?



**Speech recognition**



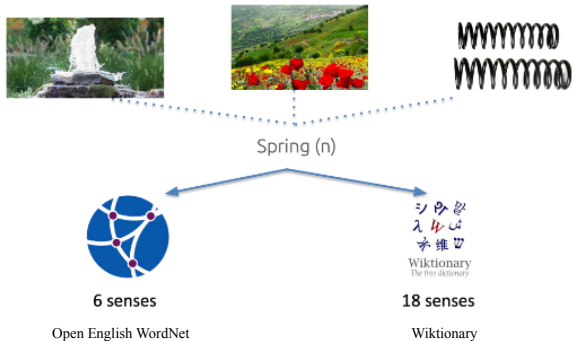
- Machine translation

The screenshot shows the Google Translate web interface. At the top, there is a hamburger menu icon followed by the word "Translate". Below this, there are three tabs: "Text" (selected), "Images", and "Websites". The source language is identified as "FRENCH - DETECTED" and the target language is "ENGLISH". The input text is "Toute faute qu'on fait est un cachot qu'on s'ouvre." and the output translation is "Any mistake we make is a dungeon we open." The interface includes a microphone icon, a speaker icon, a character count "53 / 5,000", and a "Send feedback" link at the bottom right.

# Language and Speech Technology: a few tasks

- Machine translation
- Word-sense disambiguation

In *spring*, the garden is a feast of blossom.





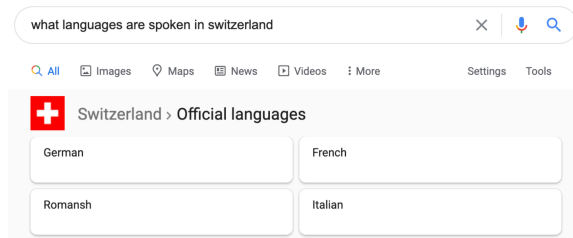
# Language and Speech Technology: a few tasks

- Machine translation
  - Word-sense disambiguation
  - Spelling error correction
- This cake is basicly sugar, butter, and flour. [→ basically]
  - We went to the store and bought new stove. [→ a new stove]
  - i'm entirely awake. [→ {I, wide}]



# Language and Speech Technology: a few tasks

- Machine translation
- Word-sense disambiguation
- Spelling error correction
- Question answering



# Language and Speech Technology: a few tasks

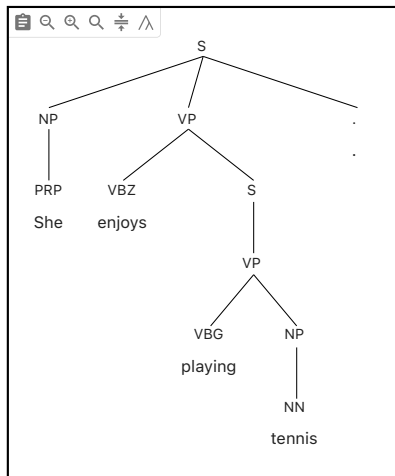
- Machine translation
- Word-sense disambiguation
- Spelling error correction
- Question answering
- Syntactic parsing

Sentence:

Berkeley Neural Parser

She enjoys playing tennis.

Parse tree:



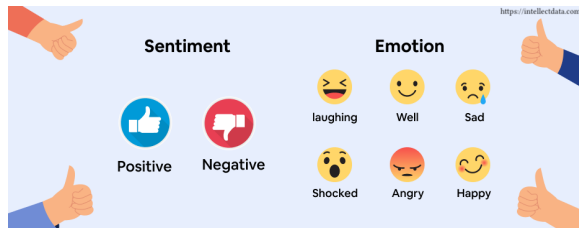
# Language and Speech Technology: a few tasks

- Machine translation
- Word-sense disambiguation
- Spelling error correction
- Question answering
- Syntactic parsing
- Text summarization



# Language and Speech Technology: a few tasks

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



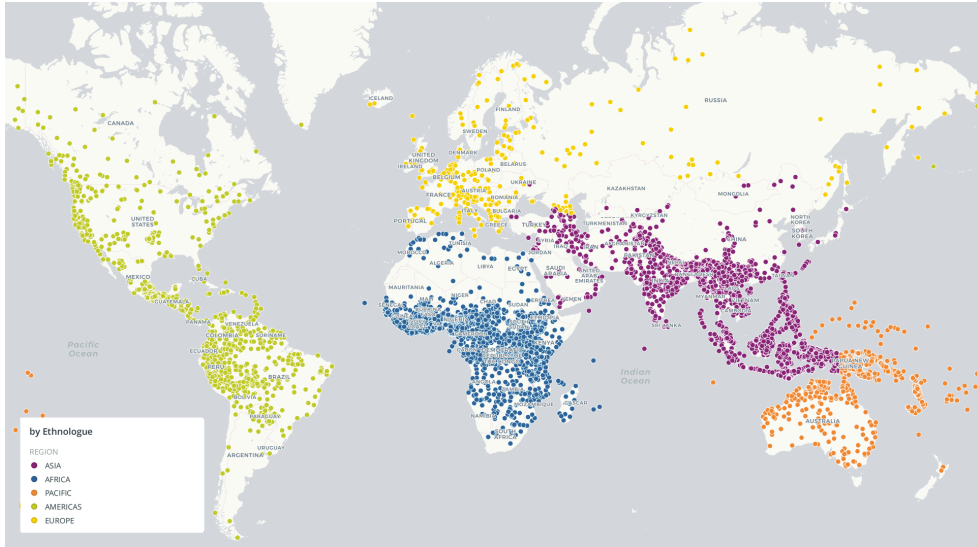
# Language and Speech Technology: a few tasks

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

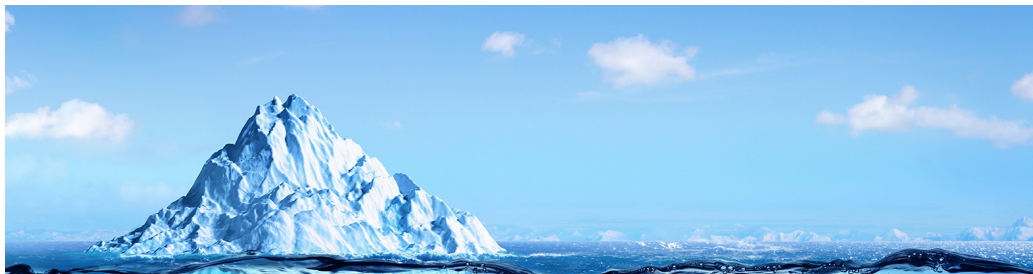


# Language and Speech Technology: Linguistic Disparity

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



# Language and Speech Technology: Linguistic Disparity







## High-resource

- Billions of documents online
- Large annotated datasets
- Large Wikipedia







**High-resource**

- Billions of documents online
- Large annotated datasets
- Large Wikipedia

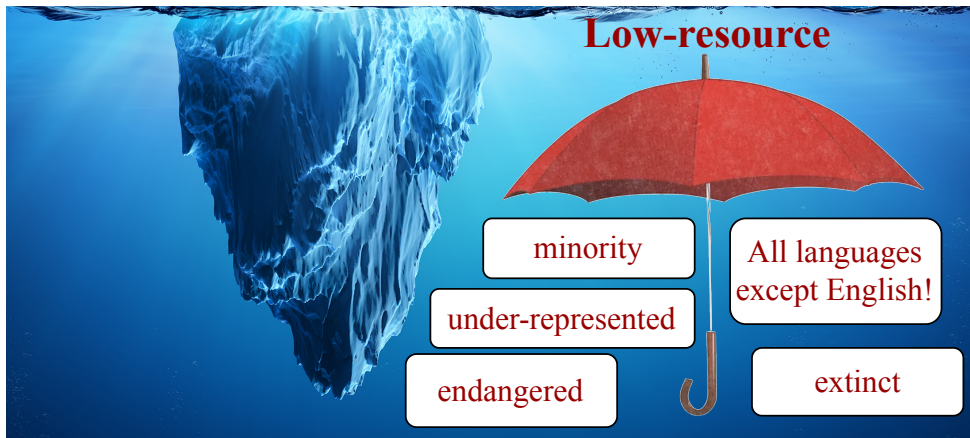
**Medium-resource**

- Millions of documents online
- Few labeled datasets
- Decent Wikipedia

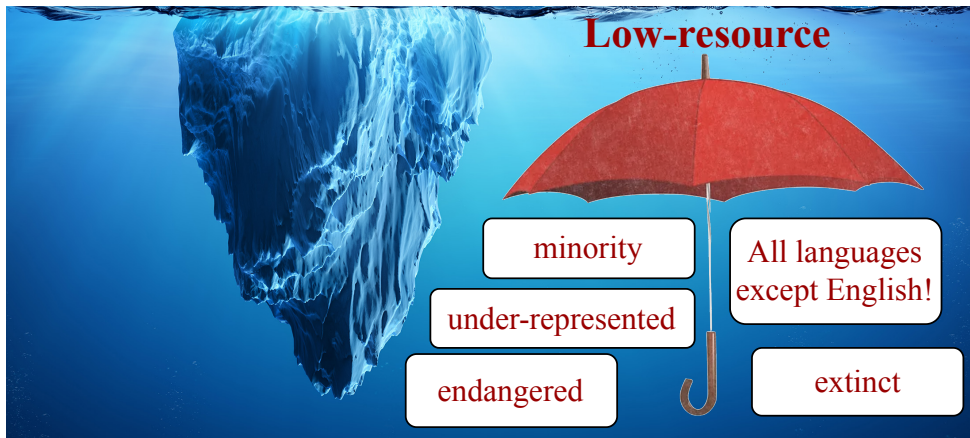
**Low-resource**

- Hundreds of documents online
- (almost) No labeled datasets
- Small Wikipedia

# Language and Speech Technology: Linguistic Disparity



# Language and Speech Technology: Linguistic Disparity



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



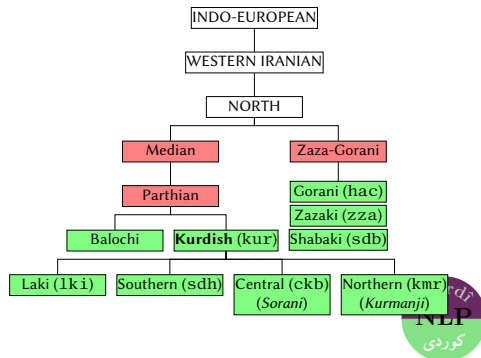
# Table of Contents

- 1 Language and Speech Technology
- 2 **Kurdish Language**
- 3 Kurdish Language Processing (KLP)
- 4 Conclusion



# Kurdish Language

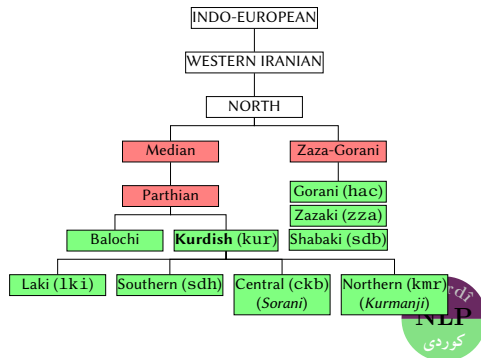
- an Indo-European language



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

# Kurdish Language

- an Indo-European language
- spoken by 20-30 million speakers

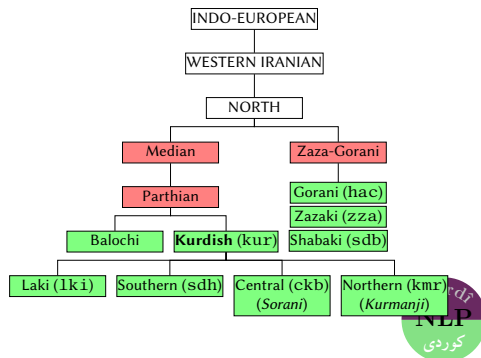


Source: <https://www.britannica.com/topic/Kurd>



# Kurdish Language

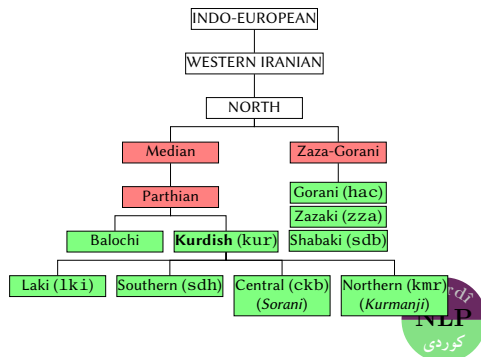
- an Indo-European language
- spoken by 20-30 million speakers
- spoken in many dialects and subdialects (*dialects or languages?*)



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

# Kurdish Language

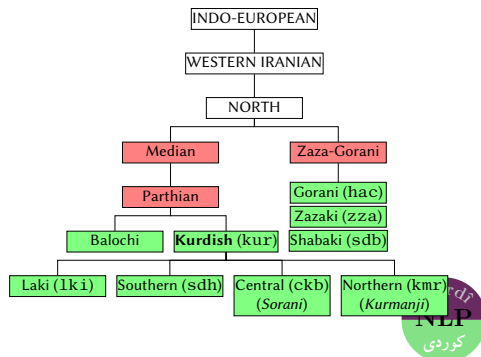
- an Indo-European language
- spoken by 20-30 million speakers
- spoken in many dialects and subdialects (*dialects or languages?*)
- has a longer oral tradition than a written one  $\Rightarrow$  *lack of data*



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

# 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

ئیمرووژ پهلونئی مەرمکە گرتوتئی خەلکێژ بئی هوول ئهژ کورونا دەوران گرتو	[lki-ar] [ra-ar]
فەلسەفە وەرچە سوڤرات، چاودێر زانستەیل سرووشتی بۆیە و کاریگەو کردار، باوەڕ، دین و ئاین خەلک نیاشتییە	[sdb-ar] [ra-ar]
وەزارەتا ئەوقافن و کاروبارین ئایینی ل هەرئێما کوردستانن ل دۆر بێهێنەدانهکا فەرمی ب هەلکەفتەکا ئایینی رۆهنکرنەک دەرکەر	[kmr-ar] [ra-ar]
له راستیدا ئەم کارمکتیرانه سهه به کۆمه‌لگای سوننه‌تی کوردستان و جیله‌کانی رابردوون	[ckb-ar] [ra-ar]
Ji ber barîna berfê li bajarê Wan û navçeya Tetwan a Bedlîsê dîmenên ciwan derketin holê.	[kmr-latn] [ckb-latn]
Bergirî lem bwareda her le yekemîn rojekanî damezrandinî komarî Turkyawe hate goşê.	[ckb-latn]

Kurdish



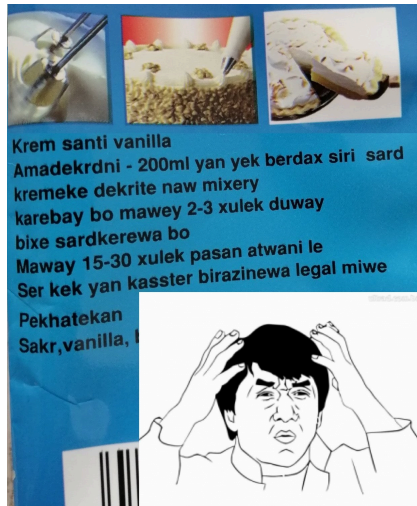
# Kurdish Language: Issues

- Too many scripts scatter readers and creates further challenges in NLP
- Written in various orthographies following different conventions



# 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*



# 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”

ئیمرووژ پهلونئی مەرمکە گرتوتئی خەلکێژ بئی هوول ئهژ کورونا دهوران گرتو	[lki-ar]
فهلسهفه ومرجه سوقرات، چاودێر زانستهيل سرووشتي بۆيه و کارێگه کردار، باوهڕ، دين و ئاين خهک نياشتيه	[sch-ar]
هزارهتا ئهوقافن وکاروبارين ئايبيني ل ههريما كوردستانن ل دۆر بيهنقهدهانهكا فهرمی ب ههلكهفتهكا ئايبيني رههنكرنهك دهركر	[kmr-ar]
له راستیدا ئهم کارهكتيرانه سهه به كومهلگای سونهتیی كوردستان و جيلهكانی رابردوون	[ckb-ar]
Ji ber barîna berfê li bajarê Wan û navçeya Tetwan a Bedlîsê dîmenên ciwan derketin holê.	[kmr-latn]
Bergirî lem bwareda her le yekemîn rojekanî damezrandinî komarî Turkyawe hate goşê.	[ckb-latn]

Kurdish



# 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”
  - *hêviya*, *hêvîya* or *hêvî ya* “hope of”?

ئیمرووژ په لاونئ مەرمکه گرتوتئ خەلکێژ بئ ههول ئهژ کورونا دهوران گرتو	[lxi-ar] [ra-ar]
فەلسەفە وەرجه سوڤرات، چاودێر زانستەیل سرووشتی بۆیه و کاریگەو کردار، باوەڕ، دین و ئاین خەلک نیاشتییە	[sch-ar] [ra-ar]
هزارهتا ئهوقافن وکاروبارین ئایینی ل هەرئما کوردستانن ل دۆر بیهنقهدهانهکا فهرمی ب ههلهگهفتهکا ئایینی رههنگرتهک دهرکر	[kmr-ar] [ra-ar]
له راستیدا ئهم کارهکتیرانه سهه به کومه لگای سونهتیی کوردستان و جیلهکانی رابردوون	[ckb-ar] [ra-ar]
Ji ber barîna berfê li bajarê Wan û navçeya Tetwan a Bedlîsê dîmenên ciwan derketin holê.	[kmr-latn] [ckb-latn]
Bergirî lem bwareda her le yekemîn rojekanî damezrandinî komarî Turkyawe hate goşê.	[ckb-latn]

Kurdish





# 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”
  - *hêviya*, *hêvîya* or *hêvî ya* “hope of”?
  - *•١٢٣٤٥٦٧٨٩*, *•١٢٣٤٥٦٧٨٩* or *0123456789*?

ئیمرووژ په لاونئی مەرمکه گرتوتئ خه لکیر بئ هوول ئهژ کورونا دهوران گرتو	[lki-ar] [ra-ar]
فهلسفه وهرجه سوقرات، چاودیر زانستیل سرووشتی بویه و کاریگه کردار، باوهر، دین و ئاین خه لک نیاشتییه	[sch-ar] [ra-ar]
هزارهتا ئهوقافن وکاروبارین ئایینی ل ههزما کوردستانن ل دۆر بیهنقه دانهکا فهرمی ب هه لکه هتهکا ئایینی رههنکرتهک دهرکر	[kmr-ar] [ra-ar]
له راستیدا ئهم کارهکتیرانه سهه به کومه لگای سونه تیی کوردستان و جیلهکانی رابردوون	[ckb-ar] [ra-ar]
Ji ber barîna berfê li bajarê Wan û navçeya Tetwan a Bedlîsê dîmenên ciwan derketin holê.	[kmr-latn] [ra-latn]
Bergirî lem bwareda her le yekemîn rojekanî damezrandinî komarî Turkyawe hate goşê.	[ckb-latn] [ra-latn]

Kurdish



# 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”
  - *hêviya*, *hêvîya* or *hêvî ya* “hope of”?
  - *•١٢٣٤٥٦٧٨٩*, *•١٢٣٤٥٦٧٨٩* or *0123456789*?
- Kurdish orthographies are phonemic, but not always:

ئیمرووژ پهلانی مەرمکه گرتوتی خه‌لکێژ بئی هوول ئەژ کورونا ده‌وران گرتو	[lxi-ar] [ra-ar]
فهلسه‌فه وهرجه سو‌قرات، چاودێر زانسته‌یل سرووشتی بۆیه و کاریگه‌ی کردار، باوه‌ر، دین و ئاین خه‌لک نیاشته‌یه	[sch-ar] [ra-ar]
ه‌زاره‌تا ئه‌وقافن وکاروبارین ئایینی ل هه‌رئما کوردستانن ل دۆر بیه‌نقه‌دانه‌کا فه‌رمی ب هه‌لکه‌فته‌کا ئایینی ره‌هنکرنه‌ک ده‌رک	[kmr-ar] [ra-ar]
له‌ راستیدا ئهم کاره‌کتیرانه سه‌ر به‌ کومه‌لگای سونه‌تی کوردستان و جیله‌کانی رابردوون	[ckb-ar] [ra-ar]
Ji ber barîna berfê li bajarê Wan û navçeya Tetwan a Bedlîsê dîmenên ciwan derketin holê.	[kmr-latn] [ra-latn]
Bergirî lem bwareda her le yekemîn rojekanî damezrandinî komarî Turkyawe hate goşê.	[ckb-latn] [ra-latn]

Kurdish



# 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”
  - *hêviya*, *hêvîya* or *hêvî ya* “hope of”?
  - *•١٢٣٤٥٦٧٨٩*, *•١٢٣٤٥٦٧٨٩* or *0123456789*?
- Kurdish orthographies are phonemic, but not always:
  - double-usage characters: *ی* for *î/y* and *و* for *u/w*

ئیمرووژ په لاونی مەرمکه گرتوتی خەلکێژ بئی هوول ئەژ کورونا دەوران گرتو	[lkt-ar] [ra-ar]
فەلسەفە وەرجه سوڤرات، چاودێر زانستەیل سرووشتی بۆیە و کاریگەو کردار، باوەڕ، دین و ئاین خەلک نیاشتییە	[sch-ar] [ra-ar]
وەزارەتا ئەوقافن و کاروبارین ئایینی ل هەرئێما کوردستانن ل دۆر بێهێنقدانەکا فەرمی ب هەلکەفتەکا ئایینی رۆهنکرێنەک دەرکەر	[kmr-ar] [ra-ar]
لە راستیدا ئەم کارمکتیرانە سەر بە کۆمەلگای سوننەتی کوردستان و جێلەکانی رابردوون	[ckb-ar] [ra-ar]
Ji ber barîna berfê li bajarê Wan û navçeya Tetwan a Bedlîsê dîmenên ciwan derketin holê.	[kmr-latn] [ra-latn]
Bergirî lem bwareda her le yekemîn rojekanî damezrandinî komarî Turkyawe hate goşê.	[ckb-latn] [ra-latn]

Kurdish



# 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”
  - *hêviya*, *hêvîya* or *hêvî ya* “hope of”?
  - *•١٢٣٤٥٦٧٨٩*, *•١٢٣٤٥٦٧٨٩* or *0123456789*?
- Kurdish orthographies are phonemic, but not always:
  - double-usage characters: *ی* for *î/y* and *و* for *u/w*
  - variations like *l*, *ll* or *†* for [†]

ئیمرووژ په لاونی مەرمکه گرتوتی خەلکێژ بئی هوول ئەژ کورونا دەوران گرتو	[lkt-ar] [ra-ar]
فەلسەفە وەرجه سوڤرات، چاودێر زانستەیل سرووشتی بۆیە و کاریگەو کردار، باوەڕ، دین و ئاین خەلک نیاشتییە	[sch-ar] [ra-ar]
وەزارەتا ئەوقافن و کاروبارین ئایینی ل هەرئێما کوردستانن ل دۆر بێهێنقەدانەکا فەرمی ب هەلکەفتەکا ئایینی رۆهنکرێنەک دەرکەر	[kmr-ar] [ra-ar]
له راستیدا ئەم کارمکتیرانه سه‌ر به کۆمه‌لگای سوننه‌تی کوردستان و جیله‌کانی رابردوون	[ckb-ar] [ra-ar]
Ji ber barîna berfê li bajarê Wan û navçeya Tetwan a Bedlîsê dîmenên ciwan derketin holê.	[kmr-latn] [ra-latn]
Bergirî lem bwareda her le yekemîn rojekanî damezrandinî komarî Turkyawe hate goşê.	[ckb-latn] [ra-latn]

Kurdish



# 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”
  - *hêviya*, *hêvîya* 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 **l**, **ll** or **†** for **[ɫ]**
  - vowel **i** missing in the Arabic-based

ئیمرووژ په لاونئی مەرمکە گرتوتئ خەلکێژ بئ ههول ئهژ کورونا دهوران گرتو	[lkt-ar] [ra-ar]
فەلسەفە وەرجه سوڤرات، چاودێر زانستەیل سرووشتی بۆیه و کاریگەو کردار، باوەڕ، دین و ئاین خەلک نیاشتییە	[sch-ar] [ra-ar]
هزارهتا ئهوقافن وکاروبارین ئایینی ل هەرئما کوردستانن ل دۆر بیهنقههادهکا فهرمی ب ههلهگهفتهکا ئایینی رههنکرتهک دهرکر	[kmr-ar] [ra-ar]
له راستیدا ئهم کارهکتیرانه سهه به کومه لگای سونهتیی کوردستان و جیلهکانی رابردوو	[ckb-ar] [ra-ar]
Ji ber barîna berfê li bajarê Wan û navçeya Tetwan a Bedlîsê dîmenên ciwan derketin holê.	[kmr-latn] [ra-latn]
Bergirî lem bwareda her le yekemîn rojekanî damezrandinî komarî Turkyawe hate goşê.	[ckb-latn] [ra-latn]

Kurdish



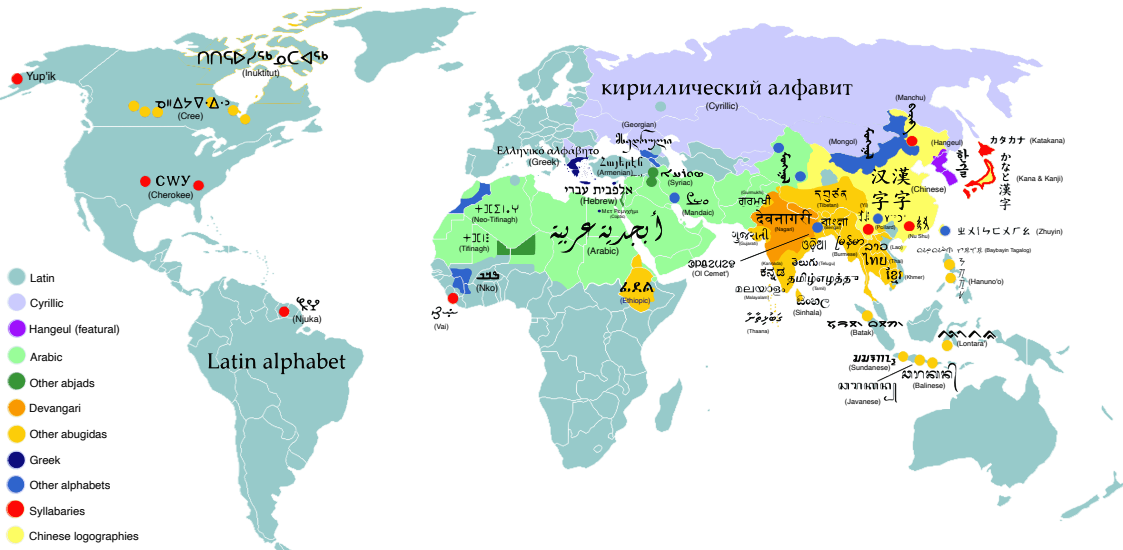
# Kurdish Language: Unconventional Writing

- Almost 300 writing systems exist (and many adopted ones)



# 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

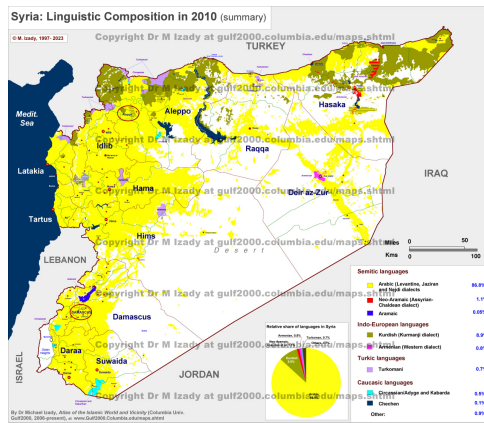




# Kurdish Language: Unconventional Writing

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

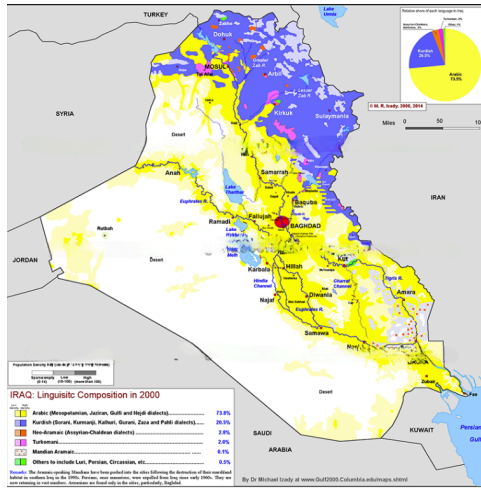
- Turkey:  
→ Turkish
- Syria:  
→ Arabic



# Kurdish Language: Unconventional Writing

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

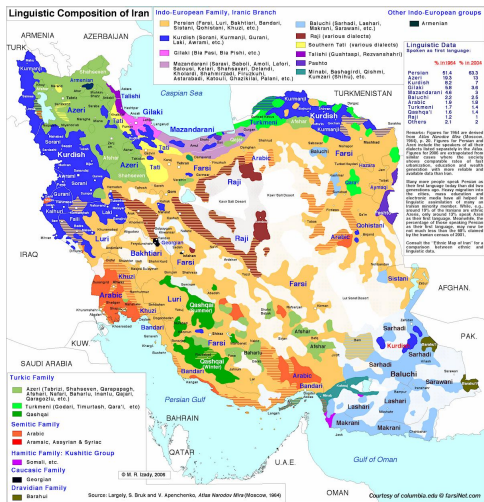
- Turkey:  
→ Turkish
- Syria:  
→ Arabic
- Iraq:  
→ Arabic and Kurdish



# Kurdish Language: Unconventional Writing

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

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



## Unconventional writing

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

- Unsystematic 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



## 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



## 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



## 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:





## 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.* (Arabic in Latin script, aka Arabizi)



## 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.* (Arabic in Latin script, aka Arabizi)
- *Tora ti na sou po* (Greek in Latin, aka Greeklish)



## 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.* (Arabic 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	یتە زون نم هیسە گە گیلکن اون جی گب زنن	یتە زوون نۆم هیسە گە گیلکۆن اون جی گب زنن
Kashmiri	Urdu	برور چھ اکھ وراسے جانور۔	برور چھ اکھ وراسے جانور۔
Kurmanji	Arabic	قایمقام الامدی بقرثوا پارزکار دھوک دا	قایمقامی ئامیدی بەرسقا پاریزگاری دھۆکی دا
Sorani	Arabic	هقر لة یەکتەم شانۆو ديارە فهدیان دەویت	هەر له یەكەم شانۆو ديارە فەهەدیان دەوێت
Sindhi	Urdu	مدیني زانهن هجرت وقت فقط هيء گھرواري سان گڈ هيئي	مدیني زانهن هجرت وقت فقط هيء گھرواري سائن گڈ هيئي



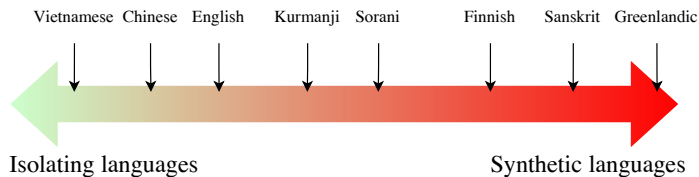
# Kurdish Language: Complex Morphology

- Kurdish has a synthetic morphology: > 2000 noun forms from a stem!



# 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 [Ahmadi et al., 2023]



POS	Morpheme per form		
	pre-stem	post-stem	average
Noun	0	3.63	3.63
Adjective	0	4.30	4.30
Verb	INTR	1.05	2.32
	TR	1.65	2.46
Average	1.35	3.1	<b>2.22</b>

Degree of synthesis in inflectional morphology of Central Kurdish based on our datasets



# 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 [Ahmadi et al., 2023]
- Complex morphotactics due to split-ergativity

0										
1										
2										
3										
4										
5										
6										
7										
8										

past stem of GIRTIN (to take, to get)  
 I got.  
 I got them.  
 I got them to/with.  
 I got them to/with again.  
 I got them also to/with again.  
 I did not get them also to/with again.  
 I was not getting them also to/with again.  
 I was not taking down them also to/with again.

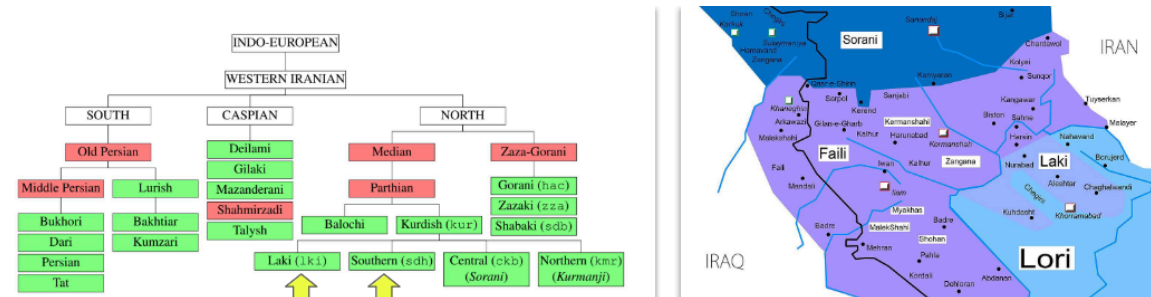
The placement of the endoclititic =îş (in green boxes) and agent marker =im (in blue boxes) with respect to the base and each other in a verb form. Note that Sorani Kurdish is a null-subject language.



# 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

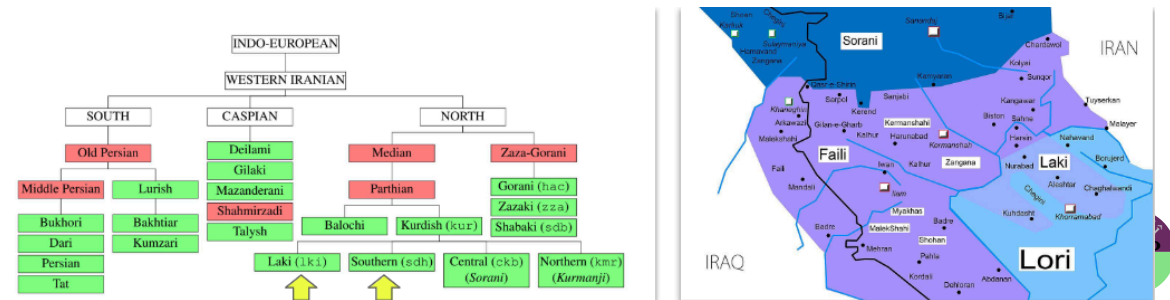




# Kurdish Language: Underrepresented Varieties

Not too many resources available for Southern Kurdish and Laki:

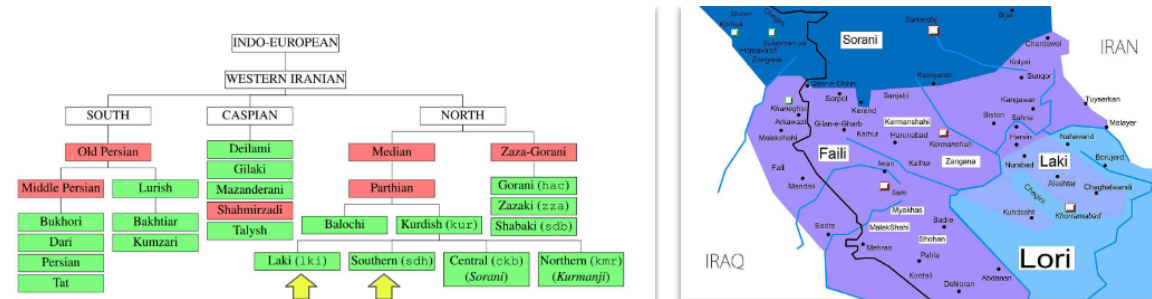
- Spoken by >2M in western Iran, border regions of Iraq, and its capital, Baghdad
- The classification of Laki is still debated



# Kurdish Language: Underrepresented Varieties

Not too many resources available for Southern Kurdish and Laki:

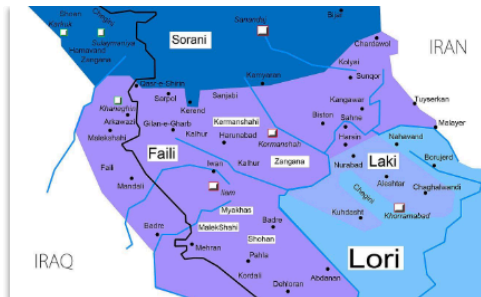
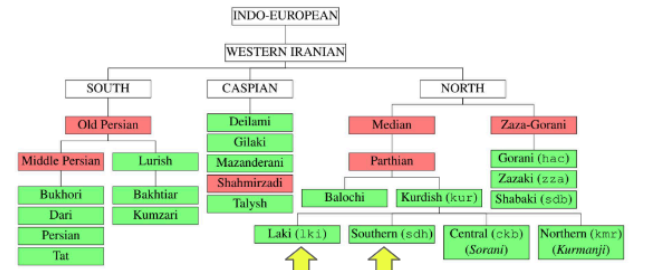
- Spoken by >2M in western Iran, border regions of Iraq, and its capital, Baghdad
- The classification of Laki is still debated
- Faced various discriminatory language policies leading to pernicious sociolinguistic effects



# Kurdish Language: Underrepresented Varieties

Not too many resources available for Southern Kurdish and Laki:

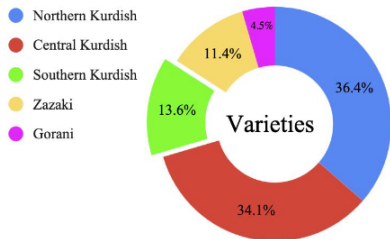
- Spoken by >2M in western Iran, border regions of Iraq, and its capital, Baghdad
- The classification of Laki is still debated
- Faced various discriminatory language policies leading to pernicious sociolinguistic effects
- Lack of children proficiency in Southern Kurdish and limited usage of the language in writing



# 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]



Percentage of the existing lexicographical resources for Kurdish



# Table of Contents

- 1 Language and Speech Technology
- 2 Kurdish Language
- 3 Kurdish Language Processing (KLP)**
- 4 Conclusion



# Kurdish Language Processing: Scientific Contributions

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

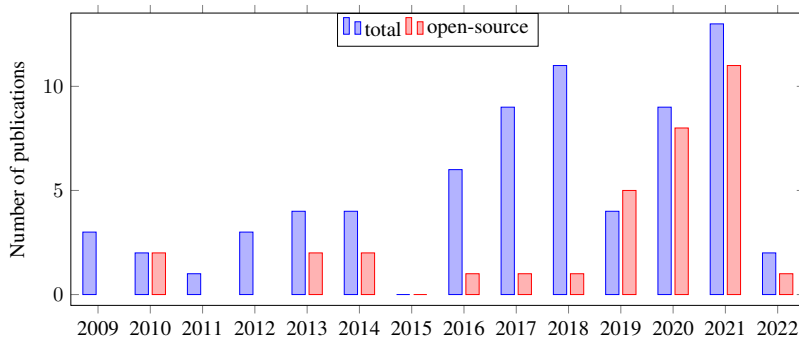
- the earliest works in the field of KLP date back to 2009 except a work in 1995 [Baban and Husein, 1995] – Why such an interval?!
- thus far, a total number of **87** publications are published in a field directly related to KLP, including non-peer-reviewed ones
- all varieties have not equally received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

## Open-source

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



# Kurdish Language Processing: Scientific Contributions

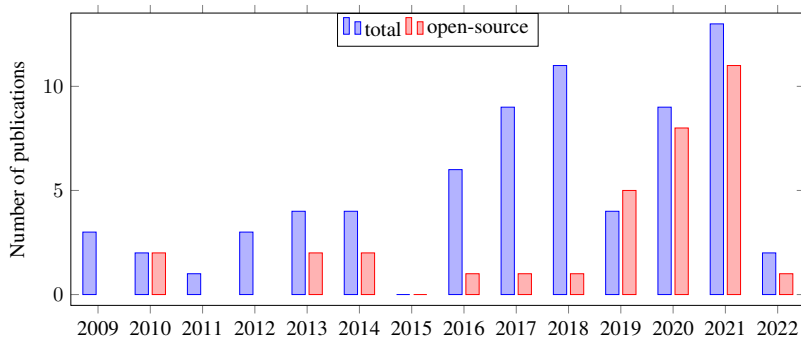


Number of scientific publications directly related to KLP per year and field

- Roughly a third provide their resources or tools under an open-source license



# Kurdish Language Processing: Scientific Contributions



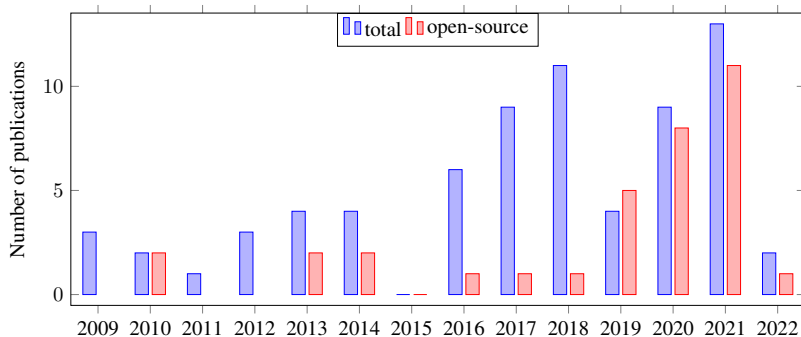
Number of scientific publications directly related to KLP per year and field

- 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





# Kurdish Language Processing: Scientific Contributions



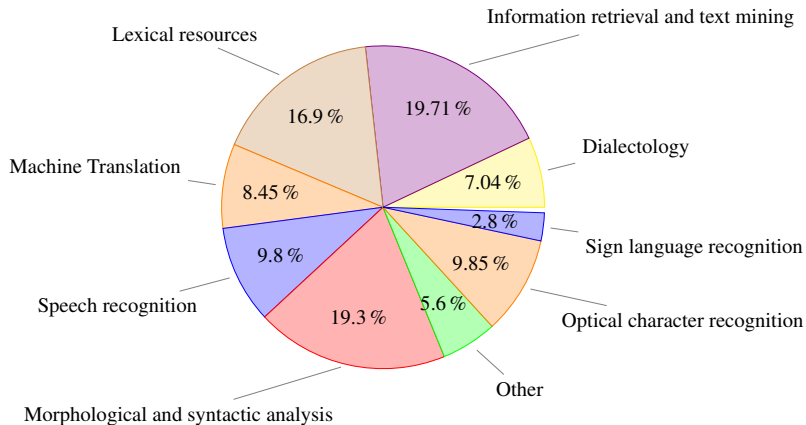
Number of scientific publications directly related to KLP per year and field

- 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:



# Kurdish Language Processing: Scientific Contributions

A range of topics have been addressed:

- Creation of lexical resources and corpora [Ahmadi et al., 2023]

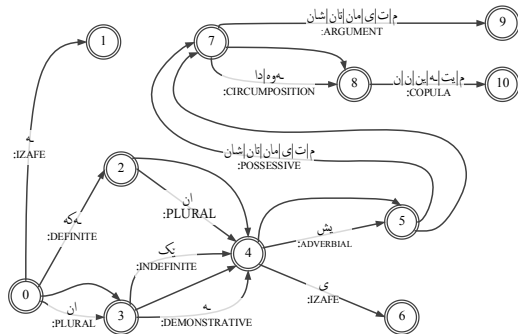
Language	639-3	Wikipedia	Common Scripts	Corpora
Northern Kurdish (Kurmanji)	kmr	ku	Latin, Central Kurdish	(Esmaili and Salavati, 2013; Ataman, 2018; Matras, 2019)
Central Kurdish (Sorani)	ckb	ckb	Central Kurdish, Latin	(Esmaili et al., 2013; Abdulrahman et al., 2019; Veisi et al., 2020; Ahmadi et al., 2020; Matras, 2019)
Southern Kurdish	sdh	-	Central Kurdish, Persian	(Fattah, 2000)
Gorani	hac	-	Central Kurdish	(Ahmadi, 2020a)
Zazaki	zza	diq	Latin	(Ahmadi, 2020a)



# 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]



A finite-state transducer for generating nouns in Central Kurdish

# 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]

Reference	Prediction	Tweet
Negative	Negative	ناخ منالی سپۆیلد و هیچ نه دیو چەن تێنەگەشتوو چەن ناشیرین چەن بێ سوود. Oh, how ugly, useless and fool (is) a spoiled and bad-mannered kid.
Neutral	Negative	یادی بە خێر 😊 بە منالی خەونەکانمان چەند گەورە بوون 🥰🥰🥰🥰 Those were the days 😊 How big our childhood dreams were 🥰🥰🥰🥰🥰
Neutral	Positive	عەشقی راستەقینە <b>وەک</b> و نوێژ وا، دواى ئەوەی نەتە هێنا نابێ سەیری دەور و بەرت بکەى. Real love is like prayer. You should not get distracted when doing it.
Negative	Negative	بە درۆی پیاوێ گەورەکان ئەلێن سیاسەت The big lies of big men (people) are called politics



# 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]

```
# Southern Kurdish
>>> model.predict("!!چەس ئمروو چە قوویمیا؟")
(('__label__sdh',), array([1.00003743]))

# Gorani
>>> model.predict("داستانێ فرمتەر و درێژتەرفنە و دەسی سەر پەیی")
(('__label__hac',), array([0.99998134]))

# Kurmanji
>>> model.predict("ئەگەر بێژم ئەز فەرھادم")
(('__label__ku',), array([0.93445575]))

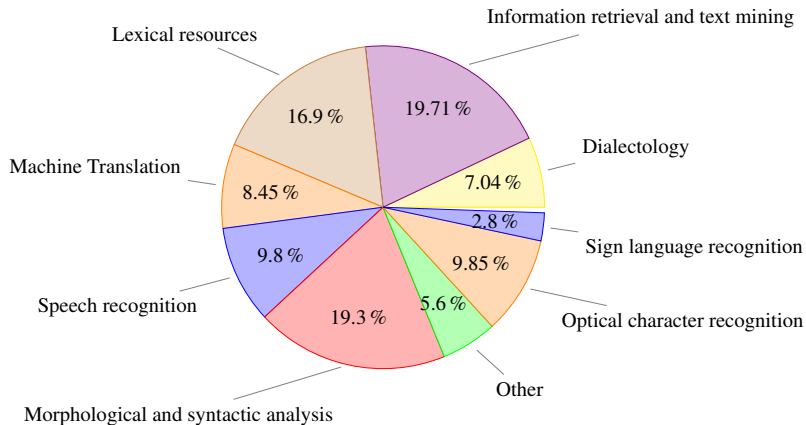
# Zazaki
>>> model.predict("Seba naye zî ganî ma rayîr û metodanê xo xurtêr bikerê.")
(('__label__zza',), array([1.00003004]))
```



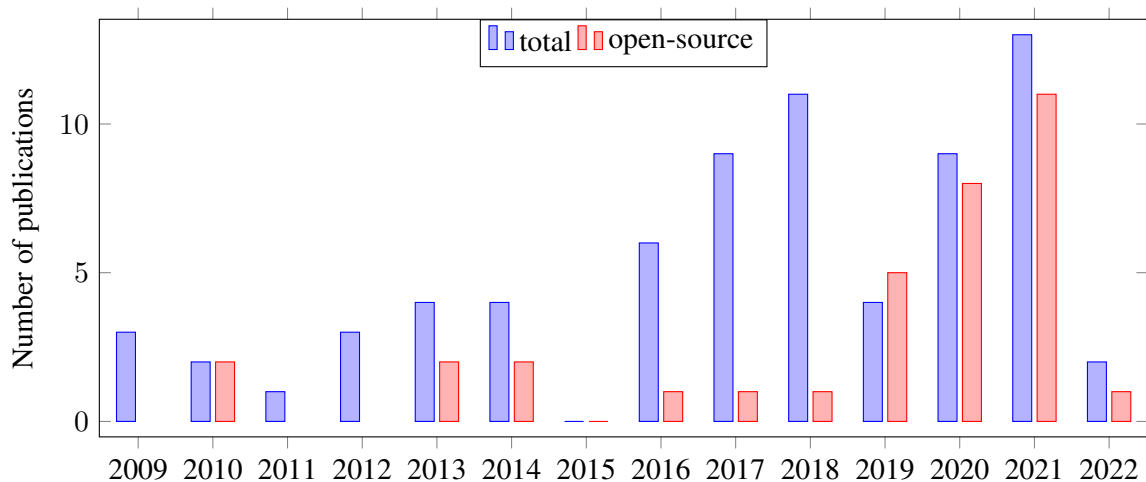
# Kurdish Language Processing: Scientific Contributions

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



# 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]



# 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



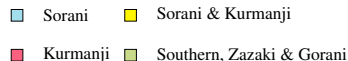
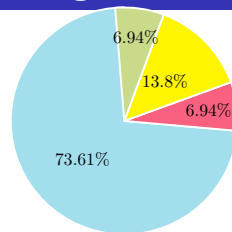
# Kurdish Language Processing: What is wrong?

- 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]



# 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



# 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



# 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 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

- Wikîferheng: <https://ku.wiktionary.org>

[wiki:far hæng]  
**Wikiferheng**  
pirzimani  
Ferhenga azad  
ویکی‌فرهنگ  
پرزیمانی  
فرهنگا ئازاد



ویکیپیدیا  
ئینسا ئۆپیدیای ئازاد



**WIKÎPEDIYA**  
Ensiklopediya azad



# Kurdish Language Processing: a few projects

- Wîkîferheng: <https://ku.wiktionary.org>
- Wîkîpediyaya kurdî: <https://ku.wikipedia.org>

[wiki:farheng]  
**Wikiferheng**  
pirzimani  
Ferhenga azad  
ویکی‌فرهنگ  
پرزیمانی  
فرهنگا ئازاد



ویکیپیدیا  
ئینسایکۆپیدیای ئازاد



**WIKÎPEDIYAYA**  
Ensiklopediya azad



# Kurdish Language Processing: a few projects

- Wîkîferheng: <https://ku.wiktionary.org>
- Wîkîpediyaya kurdî: <https://ku.wikipedia.org>
- VejînLex: <https://lex.vejin.net/en>

[wiki:farheng]  
**Wikiferheng**  
pirzimani  
Ferhenga azad  
ویکی‌فرهنگ  
پرزیمانی  
فرهنگا ئازاد



ویکیپیدیا  
ئینسایکۆپیدیای ئازاد



**WIKÎPEDIYAYA**  
Ensiklopediya azad



# 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

[wiki:farheng]  
**Wikiferheng**  
pirzimani  
Ferhenga azad  
ویکی‌فرهنگ  
پرزیمانی  
فرهنگا ئازاد



ویکیپیدیا  
ئینسا ئیکۆپیدیای ئازاد



**WIKÎPEDIYAYA**  
Ensiklopediya azad



# Kurdish Language Processing: a few projects

- Wikîferheng: <https://ku.wiktionary.org>
- Wikîpediyaya kurdî: <https://ku.wikipedia.org>
- VejînLex: <https://lex.vejin.net/en>
- Importance of Kurdish for the Tech Giants
  - Google: <https://translate.google.com>



# Kurdish Language Processing: a few projects

- Wikîferheng: <https://ku.wiktionary.org>
- Wikîpediyaya 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>

[wiki:farheng]  
**Wikiferheng**  
pirzimani  
Ferhenga azad  
ویکی‌فرهنگ  
پرزیمانی  
فرهنگه‌گا ئازاد



ویکیپیدیا  
ئینسا ئیکۆپیدیای ئازاد



**WIKÎPEDIYÂ**  
Ensiklopediya azad



# Kurdish Language Processing: a few projects

- Wikîferheng: <https://ku.wiktionary.org>
- Wikîpediyaya 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>

[wiki:farheng]  
**Wikiferheng**  
pirzimani  
Ferhenga azad  
ویکی‌فرهنگ  
پرزیمانی  
فرهنگا ئازاد



ویکیپیدیا  
ئینسایکۆپیدیای ئازاد



**WIKÎPEDIYÂ**  
Ensiklopediya azad



# Kurdish Language Processing: a few projects

- Wikîferheng: <https://ku.wiktionary.org>
- Wikîpediyaya 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

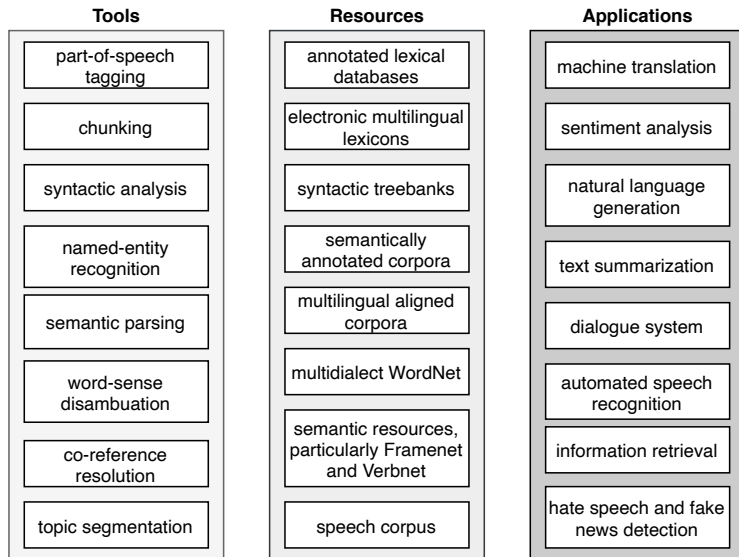
- a basic but extendable language processing toolkit
- an effort to standardize Kurdish language with all its dialects and scripts
- implemented in Python
- inspired by the functionality of relevant NLP toolkits, e.g. NLTK and spaCy
- no external NLP library is used in this toolkit
- composed of core 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>



کوردی

# KLP: Which tasks to be addressed next?



# Table of Contents

- 1 Language and Speech Technology
- 2 Kurdish Language
- 3 Kurdish Language Processing (KLP)
- 4 Conclusion



# Conclusion

- **Lessons learned:**

---

<sup>1</sup><https://en.wiktionary.org>

<sup>2</sup><https://www.wikipedia.org/>



- **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

---

<sup>1</sup><https://en.wiktionary.org>

<sup>2</sup><https://www.wikipedia.org/>

- **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

---

<sup>1</sup><https://en.wiktionary.org>

<sup>2</sup><https://www.wikipedia.org/>



- **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<sup>1</sup> and Wikipedia<sup>2</sup>

---

<sup>1</sup><https://en.wiktionary.org>

<sup>2</sup><https://www.wikipedia.org/>



- **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<sup>1</sup> and Wikipedia<sup>2</sup>
- every single user is a contributor too

---

<sup>1</sup><https://en.wiktionary.org>

<sup>2</sup><https://www.wikipedia.org/>





- **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<sup>1</sup> and Wikipedia<sup>2</sup>
- every single user is a contributor too
- **time to reconcile linguistics with computational methods for Kurdish**

---

<sup>1</sup><https://en.wiktionary.org>

<sup>2</sup><https://www.wikipedia.org/>



- **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<sup>1</sup> and Wikipedia<sup>2</sup>
- every single user is a contributor too
- **time to reconcile linguistics with computational methods for Kurdish**

- **Future directions:**

---

<sup>1</sup><https://en.wiktionary.org>

<sup>2</sup><https://www.wikipedia.org/>



- **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<sup>1</sup> and Wikipedia<sup>2</sup>
- every single user is a contributor too
- **time to reconcile linguistics with computational methods for Kurdish**

- **Future directions:**

- promote and extend technology for KLP

---

<sup>1</sup><https://en.wiktionary.org>

<sup>2</sup><https://www.wikipedia.org/>



## ● 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<sup>1</sup> and Wikipedia<sup>2</sup>
- 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

---

<sup>1</sup><https://en.wiktionary.org>

<sup>2</sup><https://www.wikipedia.org/>



## ● 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<sup>1</sup> and Wikipedia<sup>2</sup>
- 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 \$\$\$

---

<sup>1</sup><https://en.wiktionary.org>

<sup>2</sup><https://www.wikipedia.org/>

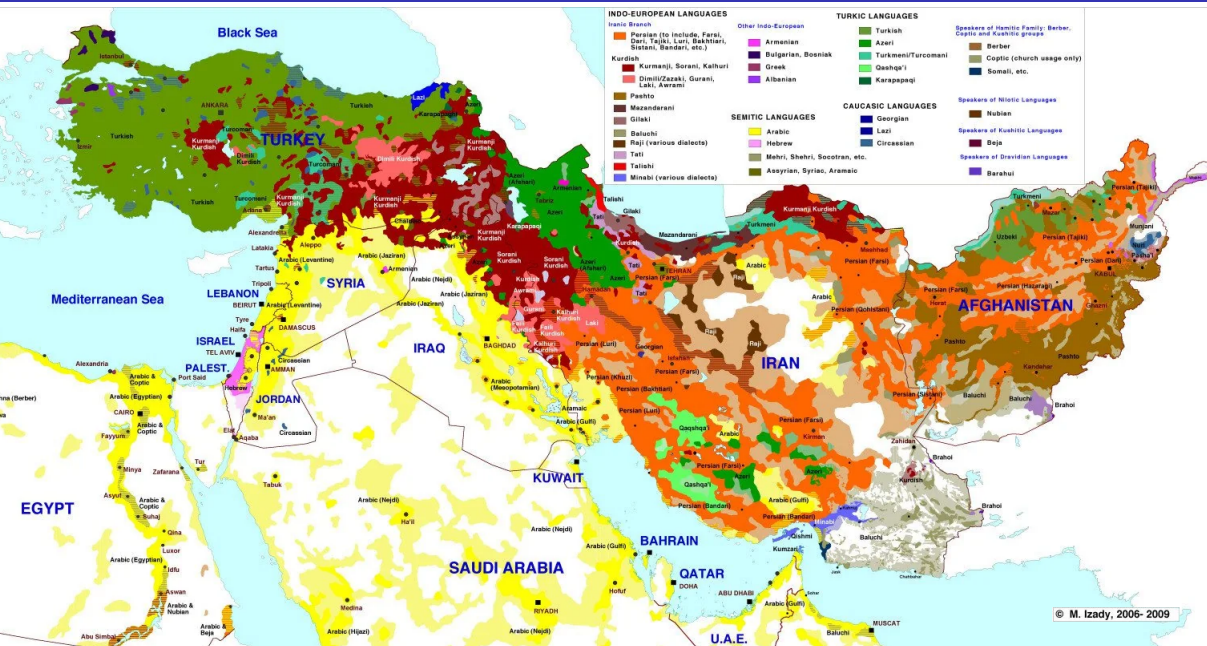


# And, the takeaway point is ...

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



# Languages of the Middle/Near East



# 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>



# 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):181–18:8.



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 Belleli 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.*



Sina Ahmadi and Aso Mahmudi (2023)

Revisiting and Amending Central Kurdish Data on UniMorph 4.0

*The 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology - ACL2023.*

