# Revisiting and Amending Central Kurdish Data on UniMorph 4.0

**Sina Ahmadi** George Mason University, VA, USA
**Aso Mahmudi** University of Melbourne, Australia

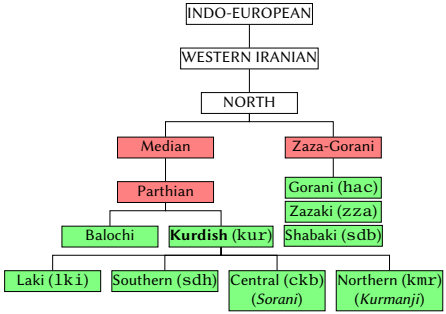
Special Interest Group on Computational Morphology and Phonology (SIGMORPHON)
Association for Computational Linguistics (ACL 2023)

July 11, 2023

# Kurdish Language

# Kurdish Language

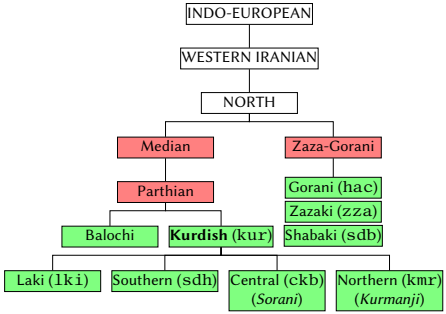
- an Indo-European language


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

# Kurdish Language

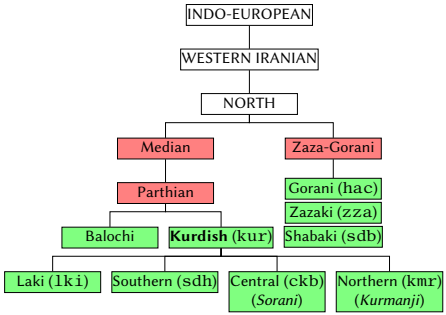
- an Indo-European language
- spoken by 20-30 million speakers


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

# Kurdish Language

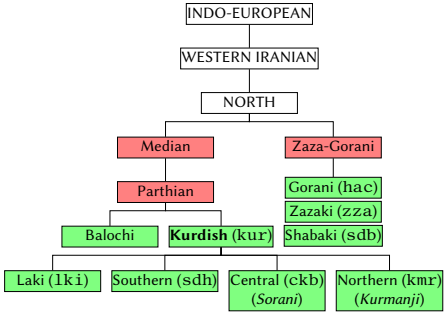
- an Indo-European language
- spoken by 20-30 million speakers
- spoken in many dialects and subdialects (*dialects* or *languages*?)


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

# Kurdish Language

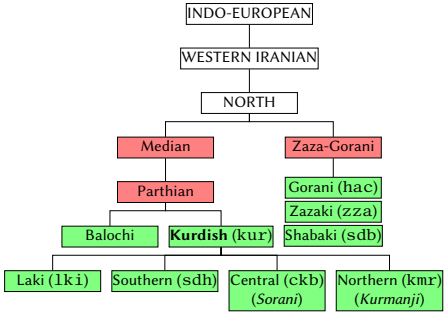
- an Indo-European language
- spoken by 20-30 million speakers
- spoken in many dialects and subdialects (*dialects* or *languages*?)
- has a longer oral tradition than a written one $\Rightarrow$ *lack of data*


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

# Kurdish Morphology

# Kurdish Morphology

1. Kurdish has a synthetic morphology → over 2000 noun forms from a stem

# Kurdish Morphology

1. Kurdish has a synthetic morphology → over 2000 noun forms from a stem
2. For verbs, it heavily relies on word-formation → simplex verb forms not exceeding 500

# Kurdish Morphology

1. Kurdish has a synthetic morphology → over 2000 noun forms from a stem
2. For verbs, it heavily relies on word-formation → simplex verb forms not exceeding 500
3. Various types of affixes including the infix *-in-* and clitics such as pronominal endoclitics

# Kurdish Morphology

1. Kurdish has a synthetic morphology → over 2000 noun forms from a stem
2. For verbs, it heavily relies on word-formation → simplex verb forms not exceeding 500
3. Various types of affixes including the infix *-in-* and clitics such as pronominal endoclitics
    - *sêv* 'apple'

# Kurdish Morphology

1. Kurdish has a synthetic morphology → over 2000 noun forms from a stem
2. For verbs, it heavily relies on word-formation → simplex verb forms not exceeding 500
3. Various types of affixes including the infix *-in-* and clitics such as pronominal endoclitics
   - *sêv* 'apple'
   - *sêv-an* 'apples'

# Kurdish Morphology

1. Kurdish has a synthetic morphology → over 2000 noun forms from a stem
2. For verbs, it heavily relies on word-formation → simplex verb forms not exceeding 500
3. Various types of affixes including the infix *-in-* and clitics such as pronominal endoclitics
    - *sêv* 'apple'
    - *sêv-an* 'apples'
    - *sêv-in-an* 'some apples'

# Kurdish Morphology

4. Complex morphotactics due to split-ergativity

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0 | | | çû | | | | past stem of ÇÛN (to go) |
| 1 | | | çû | im | | | I went |
| 2 | | | çû | im | e | | I went to |
| 3 | | | çû | im | e | ewe | I again went to (returned) |
| 4 | | de | çû | im | e | ewe | I was again going to |
| 5 | ne | de | çû | im | e | ewe | I was not again going to |

# Kurdish Morphology

## 4. 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. |

# Central Kurdish on UniMorph 4.0

# Central Kurdish on UniMorph 4.0

- UniMorph 4.0 provides a dataset for Central Kurdish 24,316 that contains word forms.

# Central Kurdish on UniMorph 4.0

- UniMorph 4.0 provides a dataset for Central Kurdish 24,316 that contains word forms.
- Initially created within the Alexina Framework [Walther and Sagot, 2010]

# Central Kurdish on UniMorph 4.0

- UniMorph 4.0 provides a dataset for Central Kurdish 24,316 that contains word forms.
- Initially created within the Alexina Framework [Walther and Sagot, 2010]
- It focuses on inflectional morphology providing paradigms of 252 lemmas (noun and verb)

# Central Kurdish on UniMorph 4.0

- UniMorph 4.0 provides a dataset for Central Kurdish 24,316 that contains word forms.
- Initially created within the Alexina Framework [Walther and Sagot, 2010]
- It focuses on inflectional morphology providing paradigms of 252 lemmas (noun and verb)
- 33 morphological features including LGSPEC1 and LGSPEC2 for *Izafe*

# Central Kurdish on UniMorph 4.0

- UniMorph 4.0 provides a dataset for Central Kurdish 24,316 that contains word forms.
- Initially created within the Alexina Framework [Walther and Sagot, 2010]
- It focuses on inflectional morphology providing paradigms of 252 lemmas (noun and verb)
- 33 morphological features including LGSPEC1 and LGSPEC2 for *Izafe*
- $<$ 1% of the word forms are assigned a unique combination of features

# Central Kurdish on UniMorph 4.0: What is wrong?

- Limited coverage of word forms and lacking diversity

| Lemma | Feature | Form in UniMorph 4.0 (Incorrect) | | Correct form | Issue |
|---|---|---|---|---|---|
| | | Original | Transliterated | | |
| *aw* 'WATER' | N;FOC | ´awš | awş | awîş ئاویش | morphophonology |
| *bûrîn* 'FORGIVE' | V;PROG;IND;SG;3;PRS;PASS | debwwrrĕĕt | debûrrêêt | debûrêt دەبوورێت | morphophonology |
| *kirdin* 'DO' | V;PROG;IND;SG;3;PRS | dekeĕ | dekeê | deka دەکا | morphophonology |
| *bezandin* 'DEFEAT (TR)' | V;PRF;SBJV;SG;1;NEG;PST | nembezandbwwayĕ | nembezandbuwayê | nembezandibuwaye نەمبەزاندبوویە | unknown morpheme *-yê* |
| *bestin* 'CLOSE (TR)' | V;PFV;SBJV;SG;1;PST | bbestmbayĕ | bibestimbayê | bimbestibaye بمبەستبایە | morphotactics |
| *kirdin* 'DO' | V;PROG;IND;PL;2;NEG;PRS;PASS | nakerĕn | nakerên | nakirên ناکرێن | missing alternation |
| *kokîn* 'COUGH' | V;IMP;SG;NEG | mekok | mekok | mekoke مەکۆکە | missing morpheme *-e* |

# Central Kurdish on UniMorph 4.0: What is wrong?

1. Limited coverage of word forms and lacking diversity
2. Unconventional writing

| Lemma | Feature | Form in UniMorph 4.0 (Incorrect) | | Correct form | Issue |
|---|---|---|---|---|---|
| | | Original | Transliterated | | |
| *aw* 'WATER' | N;FOC | ´awš | awş | awîş ئاویش | morphophonology |
| *bûrîn* 'FORGIVE' | V;PROG;IND;SG;3;PRS;PASS | debwwrrĕĕt | debûrrêêt | debûrêt دەبوورێت | morphophonology |
| *kirdin* 'DO' | V;PROG;IND;SG;3;PRS | dekeĕ | dekeê | deka دەکا | morphophonology |
| *bezandin* 'DEFEAT (TR)' | V;PRF;SBJV;SG;1;NEG;PST | nembezandbwwayĕ | nembezandbuwayê | nembezandibuwaye نەمبەزاندبوویە | unknown morpheme *-yê* |
| *bestin* 'CLOSE (TR)' | V;PFV;SBJV;SG;1;PST | bbestmbayĕ | bibestimbayê | bimbestibaye بمبەستبایە | morphotactics |
| *kirdin* 'DO' | V;PROG;IND;PL;2;NEG;PRS;PASS | nakerĕn | nakerên | nakirên ناکرێن | missing alternation |
| *kokîn* 'COUGH' | V;IMP;SG;NEG | mekok | mekok | mekoke مەکۆکە | missing morpheme *-e* |

# Central Kurdish on UniMorph 4.0: What is wrong?

1. Limited coverage of word forms and lacking diversity
2. Unconventional writing
3. Incorrect morphotactics

| Lemma | Feature | Form in UniMorph 4.0 (Incorrect) | | Correct form | Issue |
|---|---|---|---|---|---|
| | | Original | Transliterated | | |
| *aw* 'WATER' | N;FOC | ´awš | awş | awîş<br>ئاویش | morphophonology |
| *bûrîn* 'FORGIVE' | V;PROG;IND;SG;3;PRS;PASS | debwwrrĕĕt | debûrrêêt | debûrêt<br>دەبوورێت | morphophonology |
| *kirdin* 'DO' | V;PROG;IND;SG;3;PRS | dekeĕ | dekeê | deka<br>دەکا | morphophonology |
| *bezandin* 'DEFEAT (TR)' | V;PRF;SBJV;SG;1;NEG;PST | nembezandbwwayĕ | nembezandbuwayê | nembezandibuwaye<br>نەمبەزاندبووایە | unknown morpheme *-yê* |
| *bestin* 'CLOSE (TR)' | V;PFV;SBJV;SG;1;PST | bbestmbayĕ | bibestimbayê | bimbestibaye<br>بمبەستبایە | morphotactics |
| *kirdin* 'DO' | V;PROG;IND;PL;2;NEG;PRS;PASS | nakerĕn | nakerên | nakirên<br>ناکرێن | missing alternation |
| *kokîn* 'COUGH' | V;IMP;SG;NEG | mekok | mekok | mekoke<br>مەکۆکە | missing morpheme *-e* |

# Central Kurdish on UniMorph 4.0: What is wrong?

1. Limited coverage of word forms and lacking diversity
2. Unconventional writing
3. Incorrect morphotactics
4. Inaccurate morphophonological alternations

| Lemma | Feature | Form in UniMorph 4.0 (Incorrect) | | Correct form | Issue |
|---|---|---|---|---|---|
| | | Original | Transliterated | | |
| *aw* 'WATER' | N;FOC | ´awš | awş | awîş ئاویش | morphophonology |
| *bûrîn* 'FORGIVE' | V;PROG;IND;SG;3;PRS;PASS | debwwrrĕĕt | debûrrêêt | debûrêt دەبوورێت | morphophonology |
| *kirdin* 'DO' | V;PROG;IND;SG;3;PRS | dekeĕ | dekeê | deka دەکا | morphophonology |
| *bezandin* 'DEFEAT (TR)' | V;PRF;SBJV;SG;1;NEG;PST | nembezandbwwayĕ | nembezandbuwayê | nembezandibuwaye نەمبەزاندبووایە | unknown morpheme *-yê* |
| *bestin* 'CLOSE (TR)' | V;PFV;SBJV;SG;1;PST | bbestmbayĕ | bibestimbayê | bimbestibaye بمبەستبایە | morphotactics |
| *kirdin* 'DO' | V;PROG;IND;PL;2;NEG;PRS;PASS | nakerĕn | nakerên | nakirên ناکرێن | missing alternation |
| *kokîn* 'COUGH' | V;IMP;SG;NEG | mekok | mekok | mekoke مەکۆکە | missing morpheme *-e* |

# Central Kurdish on UniMorph 4.0: What is wrong?

1. Limited coverage of word forms and lacking diversity
2. Unconventional writing
3. Incorrect morphotactics
4. Inaccurate morphophonological alternations

**We estimate that 25% of the forms on UniMorph 4.0 are incorrect**

| Lemma | Feature | Form in UniMorph 4.0 (Incorrect) | | Correct form | Issue |
|---|---|---|---|---|---|
| | | Original | Transliterated | | |
| *aw* 'WATER' | N;FOC | ´awš | awş | awîş ئاویش | morphophonology |
| *bûrîn* 'FORGIVE' | V;PROG;IND;SG;3;PRS;PASS | debwwrrĕĕt | debûrrêêt | debûrêt دەبوورێت | morphophonology |
| *kirdin* 'DO' | V;PROG;IND;SG;3;PRS | dekeĕ | dekeê | deka دەکا | morphophonology |
| *bezandin* 'DEFEAT (TR)' | V;PRF;SBJV;SG;1;NEG;PST | nembezandbwwayĕ | nembezandbuwayê | nembezandibuwaye نەمبەزاندبوویە | unknown morpheme *-yê* |
| *bestin* 'CLOSE (TR)' | V;PFV;SBJV;SG;1;PST | bbestmbayĕ | bibestimbayê | bimbestibaye بمبەستبایە | morphotactics |
| *kirdin* 'DO' | V;PROG;IND;PL;2;NEG;PRS;PASS | nakerĕn | nakerên | nakirên ناکرێن | missing alternation |
| *kokîn* 'COUGH' | V;IMP;SG;NEG | mekok | mekok | mekoke مەکۆکە | missing morpheme *-e* |

# Creating a New Dataset

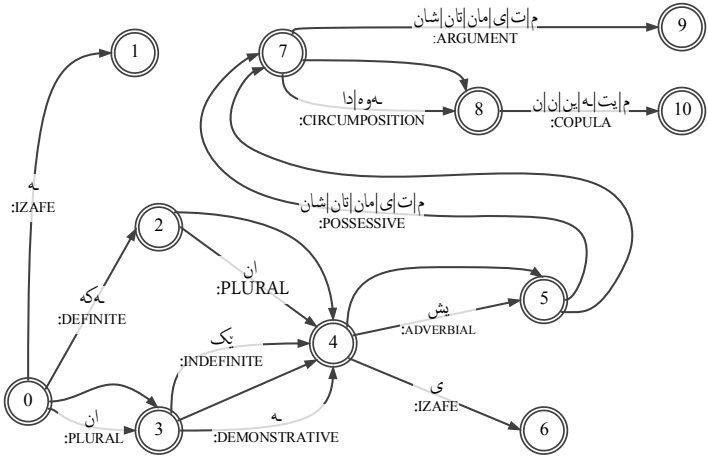
# A New Dataset for Central Kurdish

- Modeling Central Kurdish on UniMorph

| Type | Function | Ours | UniMorph |
|---|---|---|---|
| Affix | Izafe | [IZAFE] | LGSPEC1 |
| Affix | postverb adpositions | [E] [EE] | LGSPEC2 |
| Affix | postverb adverbial /ewe/ | [EWE1] | LGSPEC3 |
| Affix | disc. adpositions | [DA],[RA], [EWE2] | LGSPEC4 |
| Clitic | adverbial clitic | [ISH] | LGSPEC5 |
| Clitic | demonstrative | [DEM] | LGSPEC6 |
| Clitic | copula | [COP] | LGSPEC7 |
| Clitic | pronominal markers (argument/possessive) on transitive past verbs | [PM] | LGSPEC8 |
| Clitic | argument markers on noun/adjectives | [AM] | LGSPEC9 |

# A New Dataset for Central Kurdish

1. Modeling Central Kurdish on UniMorph
2. Finite-State Transducers



# A New Dataset for Central Kurdish

1. Modeling Central Kurdish on UniMorph
2. Finite-State Transducers
3. Morphological analysis
   analyze 1,000 random words from a corpus and manually check → **gold-standard**

# A New Dataset for Central Kurdish

1. Modeling Central Kurdish on UniMorph
2. Finite-State Transducers
3. Morphological analysis
   analyze 1,000 random words from a corpus and manually check → **gold-standard**
4. Morphological generation
   generate full paradigms (110,883 forms) for 40 lexemes → **silver-standard**

# A New Dataset for Central Kurdish

1. Modeling Central Kurdish on UniMorph
2. Finite-State Transducers
3. Morphological analysis
   analyze 1,000 random words from a corpus and manually check → **gold-standard**
4. Morphological generation
   generate full paradigms (110,883 forms) for 40 lexemes → **silver-standard**

   → Both datasets are available in two scripts of Kurdish (Latin and Arabic)

# A New Dataset for Central Kurdish

1. Modeling Central Kurdish on UniMorph
2. Finite-State Transducers
3. Morphological analysis
   analyze 1,000 random words from a corpus and manually check → **gold-standard**
4. Morphological generation
   generate full paradigms (110,883 forms) for 40 lexemes → **silver-standard**

   → Both datasets are available in two scripts of Kurdish (Latin and Arabic)

   → More diverse part-of-speech tags and lexemes

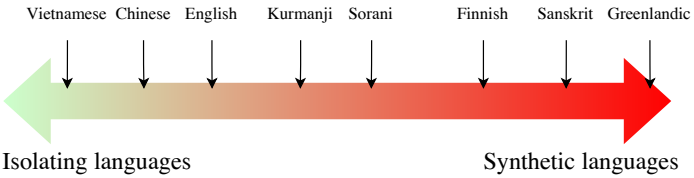
# Experiments

1. Experiments on the non-neural baseline (SIGMORPHON 2018)

| Dataset (script) | Accuracy | AED |
|---|---|---|
| UniMorph 4.0 | 48.7% | 0.97 |
| Gold-standard (L) | 63.5% | 0.99 |
| Gold-standard (A) | 67.5% | 0.88 |
| Silver-standard (L) | 61.2% | 0.98 |
| Silver-standard (A) | 65.0% | 0.75 |

# Experiments

1. Experiments on the non-neural baseline (SIGMORPHON 2018)
2. Inflectional synthesis degree [Greenberg, 1960]




| 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 | 1.68 |
| | TR | 1.65 | 2.46 | 2 |
| Average | | 1.35 | 3.1 | **2.22** |

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

# Conclusion

# Future plans

- Include a more diverse set of lexemes in the dataset

# Future plans

- Include a more diverse set of lexemes in the dataset
- Morphological variations across varieties of Kurdish

# Future plans

- Include a more diverse set of lexemes in the dataset
- Morphological variations across varieties of Kurdish
- How to deal with languages-specific features such as discontinuous morphemes?

# Future plans

- Include a more diverse set of lexemes in the dataset
- Morphological variations across varieties of Kurdish
- How to deal with languages-specific features such as discontinuous morphemes?
- Relying on this dataset, create a treebank

# Future plans

- Include a more diverse set of lexemes in the dataset
- Morphological variations across varieties of Kurdish
- How to deal with languages-specific features such as discontinuous morphemes?
- Relying on this dataset, create a treebank
- `https://github.com/unimorph/ckb`

# Thanks!

Spas!

# References


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.*

Greenberg, Joseph H (1960)
A quantitative approach to the morphological typology of language
*International journal of American linguistics, 26(3):178–194.*

Sina Ahmadi (2020)
A Formal Description of Sorani Kurdish Morphology
*https://arxiv.org/abs/2109.03942.*