A Dependency Treebank for Kurmanji Kurdish

Universal Dependencies for Kurmanji

Memduh Gökırmak & Francis Morton Tyers

Istanbul Technical University & Moscow Higher School of Economics

Kurdish - Dialects or Languages

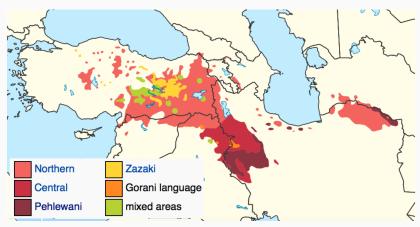
"Kurdish" traditionally refers to five dialects or languages:

- Kurmanji
- Sorani
- Zazaki
- Kelhuri
- Gorani

These groups also have subdialects, such as Suleymaniye Sorani or Behdini Kurmanji.



Areas Kurdish is Traditionally Spoken





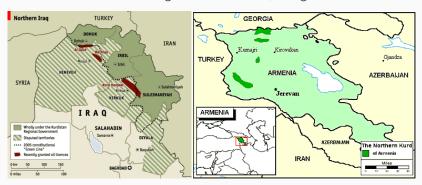
Former Official Status of Kurdish



Soviet Kurdistan existed 1923-1929/30 until it was dissolved under advice of the Foreign Ministry. Kurdish was the official language of the short-lived Republic of Mahabad (1946).

Current Official Status of Kurdish

Now official in Kurdistan Regional Government, recognized in Armenia.





Kurmanji

SAL 1 ~~ VEOSEMIN

HEJMAR 1

15GULAN 193#

Oomela Ourd) * Revue Kurde

Armanc, Awayê Xebal û Nivîsandina Hawarê

Hawar dengê zanînê ye. Zanîn xwe naxin e. xwe nasin ii me re réva felat ii xwestyn vedige Her gest go xwe has dige; digare

xwe bide nas girin. Hawara me beri ber tisti bevina zm. ané me dé hide nas girin. Lewma go 2m-

on serts heyinê a pêşîn e Hawar in pove bi her tiste uo gurdant in gurditt se bendewar e. de mijulbibe. Tine sivaget in dor e, xwe name siv-

Hawaré styaset ji civaten weloti re histiye. Bi siyaseté bila ew mijal bibin . Em ji di ware zanin, hiner n sinheté de dé bixebitin.

Awayê xebatê.- Oarenî no binare

hiçe seri, divêt je re proniviseg pet ceqirin. Me prontvisa xwe ser bingeben jerin legirive. 1 - Belavgirina Elfabéya gurdi di

nav gurdan o bingirina we. Senitandina zmanazina qurdi n bin bi bin di qomelé de belay girin ii pësdetir di sigli- gitëbë de deréxistin

2-Sehitiya zarên qurdî û berhevdanina wan. Selitti ser mirovatiya zmamė quedt digel emanên din ên ari . Sebiti ser biogeben zmane qurdi, ser dirog n awayê rabûn û pêşveqetina wî 3 - Ber hevoirina ciroq, circiroq o ber texlit laje a strapen gurdi a bireve

belaygiring wan . 4- Senitadin n belavgirina diwanin

qurdt. Bi van ve jinenigariyên şair û miroven bijarte if de ben belav girin .

5- Sehiti ser reks ii Keydeyin atranen qurdi -6- Sehiti ser her textit pizinin ou-

rdi û Qurdistanê, yên zemanê borî û xên (ro o centledins wan, Schitt ser natinen Kundistané ú nis ú sinhetén ourdí 7- Dirog n erdnigert:

Sehiti ser tevayiya diroq u erdniga riva welate Qurdistané ii ser diroga esiran, beri , pasi ii di wexto Mir-Seret de

Avayê nivîsandinê — Di hekê xmane me de heta niho geleg tist hatine eo-

tin. I i nav van gotinan de listinen rast 0 nerast ji hene. Ez qo qurd o qurdmanezman im o zmanê xwe rind dizanim û min ew bi heft hest amanen din dantye ber hev, qitqitën wi hërhënandine; keydeyën wi ji

hev derexistine, diqurim ji biyantyan betir it kenetir deboere wit bidim yorn Zmand me ird bem freh hem teng c. Freh e : Pi her liste go gurd no miint bone, dest dane wan, di wi wart de zmené gurdt hing zmenén din a ji hinan

bétir pès ve çaye, a qemiliye, a ji tu zmanen gemeli bi son de ne maye Teng e: Herçi qo ji qurdan re nenas mane a qurd pê mijal ne bune, di wi wari de zmané me rawestivave, pés ve ne cove, di cihi xwe de mave

Le zmane me ji wan zmanen e qo ber her tiste nuh, pirsen nuh diren a birive pirsin nuh ji wan çar dibin. Berwegt xelge Gelfve-Govan, bert

cardeh salan, gava tiyare ditin, tavil bizorê xwe balafir nav lê girin. Ji lewre Widely spoken yet underresourced!

Most speakers among the five

Traditionally spoken in four countries, most in Turkey

 Serious efforts to standardize since 1930s

Latin alphabet adopted in 1930s in Hawar



Orthography







Kurmanji has been written with Latin, Arabic, Cyrillic and Armenian alphabets.

Vocabulary

- Written standard very pure(-ified)
- Spoken Kurmanji often heavily influenced by prestige language
- Like Arabic, use over large geography and multiple dialects = large lexicon
- Variants for the word "thus": wisa, wilo, welê, wiha...



Grammar

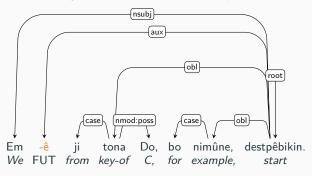
Kurmanji has a combination of interesting features that may not have existed together in a single language so far in UD, but existed separately before:

- Nouns inflect for gender, number, case and definiteness
- Split ergative, i.e. in past tenses transitive verbs agree with objects
- Future tense formed with a *clitic* after the subject
- Construct case supersedes any other case that the nominal might take
- Construct extender to add additional modifiers to a nominal



Future

Emê ji tona Do, bo nimûne, destpêbikin.



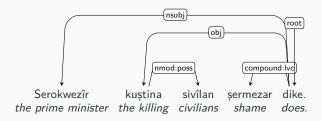
"We will begin, for example, from the key of C."



LVCs

- What is a LVC? (as far as Kurmanji is concerned):
 - Complex predicates
 - Formed of N + V
 - Common in languages Kurmanji has contact with, (Turkish, Persian)
- Conditions:
 - Another patient-like participant other than the nominal involved in the LVC
 - Nominal not inflected like simple argument to verb (i.e. nom instead of obl)
 - Can be considered secondary predication
 - Written together in the infinitive (e.g. in passive, nominal use)

Universal Dependencies

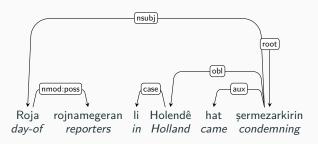


Universal Dependencies (UD) is an attempt to create a unified (as possible) framework for dependency annotation.

- Lots of freely available data!
- Cross-linguistic parsing experiments!
- Visibility for endangered, under-resourced languages



Example with the Infinitive



"Reporters' day was condemned in Holland."



The UD Kurmanji Treebank

Text	S	Т	T/S	non-proj
<i>Dr. Rweylot</i> Wikipedia	339 415	4,717 5,543	13.9 13.4	17.9 16.6
Total:	754	10,260	13.2	17.2

Table 1: Composition of the treebank. S is the number of sentences and T the number of tokens. T/S gives the average length of a sentence. The *non-proj* column gives the percentage of non-projective sentences.



Source Material

- We used works in the public domain.
- We wanted to have a corpus that wasn't restricted to a single domain
- The treebank is 50% text from a Sherlock Holmes story and 50% from Wikipedia sentences.



The Adventure of the Speckled Band



The story was published as *Dr. Rweylot* in the supplement to the famous magazine *Hawar*. Now in public domain (Syrian copyright law).

- Sherlock Holmes story translated by Bedirxan himself
- Sometimes vocabulary is old or peculiar
- Some inflections resemble the Behdini dialect
- Better for learning ergativity: conversation/narrative elements



Wikipedia



Content:

- Facts, history, biographies of e.g. Hegel or Yalçın Küçük
- Orthography:
 - More standardized than not
 - Still a good deal of variation
 - Spelling mistakes often appeared
- Process:
 - We took sentences fully covered by the analyzer
 - Randomized order of sentences



Annotation Process

```
"<Ez>"
        "ez" prn pers p1 mf sg nom @nsubi #1->2
"<Serlok>"
        "Şerlok" np ant m sg nom @root #2->0
"<Holmes>
        "Holmes" np ant m sg nom @flat #3->2
"<im>"
        "bûn" vbcop pri p1 sg @cop #4->2
"<:>"
        ";" sent @punct #5->2
"<ev>"
        "ev" prn dem mf sp nom @nsubj #6->8
"<iî>"
        "iî" emph @discourse #7->8
"<hevalê>"
        "heval" n m sg con def @parataxis #8->2
"<min>"
        "ez" prn pers p1 mf sg obl @nmod:poss #9->8
"<6>"
        "yê" con m sg @dep #10->8
"<0717>"
        "ezîz" np ant m sg obl @nmod:poss #11->8
"<Wetsin>"
        "Wetsin" np ant m sg nom @appos #12->8
"<0>"
        "bûn" vbcop pri p3 sg @cop #13->8
11 < >11
        "." sent @punct #14->2
```

- Annotated in text editors
- Annotation tools bad at fixing tokenization
- Text was run through an analyzer and a rule-based disambiguator
- Disambiguation verified and dependencies annotated in vim



Kurmanjî Analyzer

```
$ echo Tu wek çêlekê şer dikî | apertium -d . kmr-debug
"<Tu>"
        "tu" prn pers p2 mf sg nom SELECT:197
        "tû" n f sg nom def SELECT:197
        "tû" n f pl nom def SELECT:197
"<wek>"
        "wek" pr @case MAP:304
"<cêlekê>"
        "çêl" n f sg obl ind @nmod MAP:302
        "çêl" n f sg con ind @nmod MAP:302
        "çêlek" n f sg obl def @nmod MAP:302
        "cêlek" n f sg obl dem REMOVE: 154
        "cêlek" n f sg voc def REMOVE:171
"<ser>"
        "ser" n m sg obl def @dobj SELECT:134 MAP:308
        "şêr" n m sg obl def @dobj SELECT:134 MAP:308
        "ser" n m pl nom def SELECT: 134
        "sêr" n m pl nom def SELECT: 134
        "sêr" n f pl nom def SELECT: 134
        "ser" n m sg nom def SELECT:134 REMOVE:179
        "şêr" n m sg nom def SELECT:134 REMOVE:179
        "şêr" n f sg nom def SELECT:134 REMOVE:179
"<dikî>"
        "kirin" vblex tv pri p2 sg
        "dik" n m sg obl dem REMOVE:154
        "dîk" n m sg obl dem REMOVE:154
"<.>"
        "." sent @punct MAP:310
```

- Developed as part of GSoC 2016 project
- 85% coverage of Wikipedia, higher on news corpora
- Robust to dialect variation
- Disambiguation with Constraint Grammar (CG)
- ullet \sim 100 disambiguation rules



Evaluation

We evaluated the data using **tenfold** validation. In all models we used **word2vec** trained on Wikipedia.

The analysis was done with:

- Just UDpipe
- UDpipe with a dictionary of all forms from the analyzer

The parsing was done using:

- UDpipe
- BiST
- Maltparser



Parsing Results

Parser	UAS [range]	LAS [range]		
Maltparser	69.4 [64.5, 76.7]	61.5 [57.3, 65.3]		
BiST	71.2 [68.1, 74.4]	63.8 [60.7, 67.5]		
UDpipe	73.1 [66.9, 77.6]	65.9 [59.6, 68.3]		
Maltparser [+dict]	71.2 [67.8, 78.7]	64.0 [60.8, 69.3]		
BiST [+dict]	72.7 [69.4, 74.5]	66.3 [63.7, 68.5]		
UDpipe [+dict]	74.3 [72.6, 77.2]	67.9 [65.6, 70.1]		



POS Tagging & Features

System	Lemma	POS [range]	Morph [range]
UDpipe	88.3 [85.3, 89.6]	88.2 [85.5, 90.8]	78.6 [75.4, 80.1]
UDpipe [+dict]	94.6 [93.9, 95.7]	93 [91.8, 93.8]	85.9 [84.2, 87.6]



CoNLL Shared Task

Parser	LAS	UAS	
C2L2 (Ithaca)	47.53	54.51	
IMS (Stuttgart)	46.7	54.73	
HIT-SCIR (Harbin)	44.7	52.55	
Koç University (İstanbul)	42.11	49.32	
LATTICE (Paris)	41.71	51.8	
UALING (Tucson)	40.57	48.66	
UParse (Edinburgh)	39.76	52.22	
Orange – Deskiñ (Lannion)	38.31	50.76	
LIMSI-LIPN (Paris)	35.59	48.67	
Stanford (Stanford)	35.05	47.71	
OpenU NLP Lab (Ra'anana)	34.94	51.24	
ParisNLP (Paris)	34.8	47.76	
darc (Tübingen)	33.06	46.32	** 10
ÚFAL – UDPipe 1.2 (Praha)	32.89	46.3 ®	** Apertium
BASELINE UDPipe 1.1 (Praha)	32.35	46.2	
			22

Future Work

- Expanding the treebank
- More strictly standard, modern data, possibly news texts
- Fixes and maintenance
- Work on other Kurdish languages, cross-linguistic experiments

