A Dependency Treebank for Kurmanji Kurdish

Memduh Gökırmak
Department of Computer Engineering
Istanbul Technical University
Turkey
gokirmak@itu.edu.tr

Francis M. Tyers
School of Linguistics
Higher School of Economics
Russia
francis.tyers@uit.no

Abstract

This paper describes the development of the first syntactically annotated corpus of Kurmanji Kurdish. The corpus was used as one of the *surprise* languages in the 2017 CoNLL shared task on parsing Universal Dependencies. In the paper we describe how the corpus was prepared, some Kurmanji specific constructions that required special treatment, and we give results for parsing Kurdish using two popular data-driven parsers.

1 Introduction

With current end-to-end pipelines for tokenisation, tagging and parsing, such as UDPipe (Straka et al., 2016), a treebank is no longer simply a collection of annotated sentences, but could be considered a vital basic language resource. Given just the treebank a statistical model can be trained which performs everything up to dependency parsing.

This paper describes such a treebank for Kurmanji Kurdish, a language spoken in parts of Iran, Iraq, Syria, Armenia and Turkey. The treebank was created as one of the *surprise languages* for the CoNLL 2017 shared task in dependency parsing (Zeman et al., 2017);[1] but it is hoped that it provides a template for further development of language technology for Kurmanji.

The paper is laid out as follows, in Section 2 we give a brief sociolinguistic and typological overview of the Kurdish. Then in Section 3 we describe some prior work on computational resources and tools for Kurmanji. In Section 4 we describe the composition of the corpus, and in Section 5 we describe some details of the annotation guidelines, paying attention to Kurmanji-specific phenomena. Section 6 reports on a small experiment with three popular data-driven parsers, and is followed by some avenues for future work in Section 7 and conclusions in Section 8.

2 Kurdish

Kurmanji Kurdish (also referred to in the literature as 'Northern Kurdish') is an Indo-Iranian language spoken by approximately 14 million people throughout the Middle East. It is a recognised minority language in Armenia (Simons and Fennig, 2017). Kurmanji over the past century has become the most prominent Kurdish language, partly due to the fact that its speakers are a majority among speakers of Kurdish languages, and partly due to intense cultural and political activity centered around the Kurmanji language. Manuscripts in what could be considered a precursor to Kurmanji have been discovered from five centuries back or more, but the most intense efforts in the creation of a literary written standard of Kurmanji were in the 1920s and 30s onward throughout the 20th century. Through the work of writers, academics and intellectuals like Celadet Bedirxan and his colleagues at *Hawar*, the Damascene Kurdish magazine where the Latin Kurdish alphabet was first adopted, Kurmanji has accumulated a respectable literature and a standard register has been created. Despite all of this activity and possibly due to the 'prestige' status of other languages in the region,[2] many speakers of the various dialects of Kurmanji are not aware of a Kurdish literature, and some are even shocked to learn that Kurdish languages are written at all.

Kurmanji has two grammatical genders, masculine and feminine; four cases: nominative, oblique, construct and vocative; and definiteness marked on nouns. The language has prepositions and postpositions, and also combinations of these which form circumpositions. Verbs are formed from two stems, past and present.

[1] http://universaldependencies.org/conll17/

[2] Such as Arabic, Persian and Turkish

Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), pages 64-72,
Pisa, Italy, September 18-20 2017

Figure 1: The Kurmanji speaking area (dark grey) within the wider Kurdish speaking area (light grey). The areas where Kurmanji is most widely spoken straddle the borders of Iran, Iraq, Syria and Turkey.

Regarding syntax, the language is primarily subject–object–verb, with auxiliaries following the main verb and split-ergative alignment, where past-tense transitive verbs agree with the person and number of the syntactic object rather than the subject. Noun phrases are largely head initial, with modifiers following the head noun, exceptions to this are determiners and numerals which precede the modified noun. The language has a fairly strict constituent order, and the morphology is of the fusional type with the complexity being similar to that of Icelandic.

3 Prior work

There are a number of reference grammars of Kurmanji available, the most widely-known being Thackston (2006). We also made use of the grammar by Bedirxan and Lescot (1990), and consulted the grammar by Aktuğ (2013). Many other grammars are available, including several different writings by Celadet Ali Bedirxan himself, in most languages of the Middle East, French and English. Many of these grammars are written for the purpose of teaching beginners, and most of these introductory grammars lack important details required for proper linguistic reference. Many grammars also have a good deal of influence from majority languages in the countries they were written. This particularly comes to light when the writer of a grammar describes and thinks about elements of Kurmanji with analogy to Turkish.

A text corpus of Kurmanji and Sorani Kurdish by the name of *Pewan* was introduced in Esmaili and Salavati (2013). Pewan is a plaintext corpus created for the purpose of information retrieval, and was the first publically-available digital corpus of Kurdish. The corpus is unfortunately not freely available, being based on texts under restrictive copyright provisions.

Another lexical resource for Kurdish, although again unfortunately not freely available, is KurdNet (Aliabadi et al., 2014). This is an effort to build a WordNet-like resource for all variants of Kurdish, including Kurmanji.

Walther et al. (2010) describes the rapid development of a morphological analyser and part-of-speech tagger for Kurmanji based on a raw corpus and Thackston's reference grammar (Thackston, 2006). They start by defining part-of-speech and morphological categories, and then build a morphological description of Kurmanji in their formalism. They train a maximum-entropy based tagger using a number of different unsupervised methods achieving an accuracy of 85.7% on a hand-tagged evaluation corpus of thirteen sentences. The semi-automatically created lexicon described was released under a free/open-source licence allowing it to be incorporated, after improvement in the Apertium morphological analyser for Kurmanji (see §4.2).

4 Corpus

The corpus comprises of text from two domains, the first is a short Sherlock Holmes story, *Dr. Rweylot*,[3] which was translated into Kurmanji by Segman (1944) and published in the *Rohanî* journal in Damascus.

The motivation behind choosing a story text as opposed to news text was threefold. First of all being published in 1944 by an author who died in 1951,[4] the text is out of copyright. Secondly, having a whole story annotated as opposed to individual sentences will be interesting when looking at problems such as co-reference resolution. Finally, the orthography is close enough to the modern orthography that any differences can be easily handled.

The text was available through the Kurdish Digital Library of the Paris Kurdish Institute[5] as a PDF file. The PDF had already been processed with an OCR system, and the resulting body of text was accurate enough to use with some manual fixing of errors resulting from the OCR process.

[3]Original title: *The Adventure of the Speckled Band*.

[4]Bişarê Segman is widely believed to be a nom de plume of Celadet Berdixan, who died in 1951.

[5]`http://bnk.institutkurde.org/`

Text	S	T	T/S	*non-proj*
Dr. Rweylot	339	4,717	13.9	17.9
Wikipedia	415	5,543	13.4	16.6
Total:	780	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.

The remainder of the treebank is made up of sentences selected randomly from the Kurdish Wikipedia.[6] From the randomly-selected sentences, we excluded those which were not in Kurmanji, those with too many orthographic errors and, for legal reasons, those dealing with topics considered *controversial* in Turkey.

4.1 Orthography

Kurmanji Kurdish, unlike Sorani Kurdish, is primarily written using the Latin script, rather than the Perso-Arabic script, ever since *Hawar* adopted the Latin script in the 1930s. Both, however, use *alphabets* as their primary writing system: Sorani uses a modified version of the Perso-Arabic abugida, by introducing mandatory vowels. Kurmanji's alphabet includes several letters with diacritics: circumflexes to mark long vowels, and cedillas to mark palato-alveolar affricates ş /ʃ/ and sibilants ç /tʃ/. The script was also devised by Celadet Bedirxan.

In both the Sherlock Holmes story and the Wikipedia sentences, the orthography was not standardised. This is an issue in written Kurmanji, where many can more or less write in a certain *literary* dialect but few will produce texts that overlap completely in terms of orthography. Depending on the writer's dialect, the word *ku* 'that' might be written *ko*, *heye* 'there is' might be written as *heya*, adpositions might have slight variations and spelling may vary to represent the differences in pronunciation. In order to be able to represent this variety in the treebank we have maintained the differently spelled words in the form column of the CoNLL-U file,[7] and used the variants that exist in the morphological analyser in the lemma column, e.g. both *heya* and *heye* will have the lemma *hebûn* (the existential copula).

Another orthography issue becomes apparent in tokenisation. In the Sherlock story, in some cases negation is written analytically where it would be synthetic in a more modern text. Example (1a) shows negation written separately from the verb, while in example (1b) it is written together.[8]

(1) a. *Zimanê* *wê* ***ne***
Tongue-CON she-OBL NEG
digeriya.
turn-PROG.NARR.2SG
'Her tongue was not turning.'

b. *Zimanê* *wê*
Tongue-CON she-OBL
ne*digeriya.*
NEG-turn-PROG.NARR.2SG
'Her tongue was not turning.'

We have kept this syntactic variety as it is likely that many sentences parsed with any system based on this treebank will also have some non-standard syntactic elements, and standardising and fixing too much may lead to a less robust system.

Throughout the paper, we use { and } symbols to mark where contraction has taken place in the dependency trees, for example *Ezê* 'I will' will be shown as {Ez- -ê}. contracted with the first person singular pronoun.

4.2 Preprocessing

Preprocessing the corpus consists of running the text through the Kurmanji morphological analyser[9] available from Apertium (Forcada et al., 2011), which also performs tokenisation of multi-word units based on the longest match left-to-right. The morphological analyser returns all the possible morphological analyses for each word based on a lexicon of around 13,800 lexemes. After tokenisation and morphological analysis, the text is processed with a constraint-grammar (Bick and Didriksen, 2015) based disambiguator for Kurmanji consisting of 85 rules which remove inappropriate analyses in

[6] Database dump: `kuwiki-20150901-pages-articles.xml.bz2`

[7] CoNLL-U is the file format used in Universal Dependencies for storing treebanks. A description of the format can be found here: `http://universaldependencies.org/format.html`

[8] The tags used in the glosses are: CON = construct case, OBL = oblique case, PROG = progressive aspect, NARR = narrative tense, 2SG = second person singular.

[9] `https://svn.code.sf.net/p/apertium/svn/languages/apertium-kmr`

context. For example, there is a systematic ambiguity between the past participle and the second-person singular past tense of the verb. One rule removes the participle reading if there is no following auxiliary verb. Applying these rules reduces the average number of analyses per word from around 2.87 to around 1.47.

4.3 Formats

The native format of the treebank is the VISL format (Bick and Didriksen, 2015). This is a text-based format where surface tokens are on one line, followed by analyses on the subsequent line. The reason for choosing this format was that it was more convenient for hand-annotation, and was the format that the morphological analyser and constraint grammar output. In Appendix A we present, for reference, a sentence in VISL format.

4.3.1 CoNLL-U

In order to convert to the standard CoNLL-U format, we needed to do some additional processing:

- The morphological analyser sometimes tokenises two space-separated tokens into a single token, for example *li ber* 'in front of' is tokenised as a single token. When the surface form and the lemma had an equal number of spaces were split into multiple tokens.
- Parts of speech and morphological features were converted from Apertium standard to Universal Dependencies using a lookup table and set longest-overlap algorithm.
- In multiword tokens where there is a single surface form with multiple syntactic words, the sub-word tokens are created using a language-independent longest-common-subsequence algorithm with the surface form and the underlying lemma. For example, LCS(*ezê*, *ez*) = *ez* and LCS(*ezê*, *dê*) = ê.
- The special `SpaceAfter=No` feature, used in training tokenisers, was added automatically to the MISC column of CoNLL-U by a script.

After these transformations a valid CoNLL-U file is produced which can be used in training most popular statistical parsers.

5 Annotation guidelines

The annotation guidelines are based on Universal Dependencies (Nivre et al., 2016), an international collaborative project to make cross-linguistically consistent treebanks available for a wide variety of languages. The Kurmanji treebank is based on version 2.0 of the guidelines which were published in December, 2016.

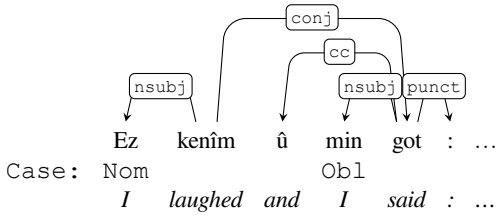
We chose the UD scheme for the annotation as it provides ready-made recommendations on which to base annotation guidelines. This reduces the amount of time needed to develop bespoke annotation guidelines for a given language as where the existing *universal* guidelines are adequate they can be imported wholesale into the language-specific guidelines.

In the following subsections we describe some particular features of Kurmanji that are interesting or novel with respect to the Universal Dependencies annotation scheme.

5.1 Alignment

Kurmanji, like other Kurdish languages, is split ergative. This is similar to the languages of the (relatively) closely related Indo-Aryan family. Ergativity does not, however, exist in most other Indo-Iranian languages. With intransitive clauses and in non-past-tense transitive clauses, the verb agrees with the most agent-like argument (typically in nominative case). However in past-tense transitive clauses, the verb agree with the most patient-like argument, which is usually in nominative case, while the most agent-like is in the oblique case. This is different to the Indo-Aryan system, which primarily uses *aspect*, rather than tense, to assign ergativity.

The following sentence in the treebank provides a good example of the contrast between transitive and intransitive sentences in the past tense: *Ez kenîm û min got:* … 'I laughed and I said: …'



Note the intransitive verb *kenîm* 'laughed' has the subject in nominative, while the transitive verb *got* 'said' has the subject in the oblique.

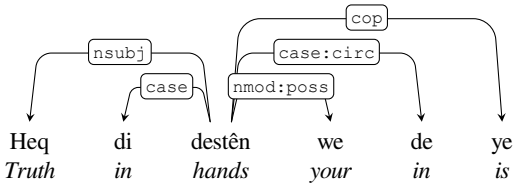
5.2 Contracted prepositions

Similar to the preposition–pronoun combinations in the Celtic languages, and like the Spanish *contigo* 'with you', Kurmanji has four prepositions which contract with third-person singular complements.

These are *bi* ‘with’, *ji* ‘from’, *di* ‘at/in’ and *li* ‘at/in’. They are dealt with in the annotation by assigning to syntactic words to the surface form, one representing the preposition and the other representing the pronoun.

5.3 Circumpositions

In addition to prepositions, Kurmanji also employs circumpositions, where a preposition and a postposition encircle the same noun phrase. In some situations, both the preposition and postposition *must* appear together, e.g. *di … de* ‘in …’. In other situations the prepositions can be used on their own. In the latter situation the postposition either modifies or gives a more nuanced meaning to and thus refines the meaning of the preposition. Consider the following example, *Heq di destên we de ye.* ‘The truth is in your hands’.



Either the preposition or the postposition can be elided, this phenomenon occurs more frequently in colloquial speech. The elided adposition is the *non-essential* one. If a postposition is part of a circumposition, we annotate it with the language-specific relation `case:circ`.

5.4 Construct case

The construct case in Kurdish is used to link a head noun to adjectival or nominal modifiers.

Construct inflection on the head noun signifies that the following word modifies the initial word. When more than one word modifies the initial word in a construct structure, a *construct extender* is used to show that the second modifier also modifies the initial noun, as opposed to modifying the last noun in the noun–noun structure.

If the phrase only has two elements, then sometimes the construct inflection can be dropped. In this case the head noun is inflected in the nominative case.

The construct case *overrides* any other inflection that the noun might have if it were not in a construct phrase. See Figure 2 for an example of how the construct inflection overrides the inflection from verbal subcategorisation.

5.5 “Light” verbs

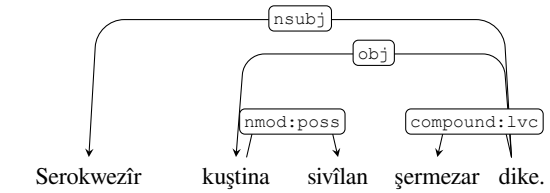
We use the term *light verb* to refer to the complex predicates formed of a nominal plus a verb which is used as a single predicate. These are common in languages that Kurmanji is in close contact with, such as Persian and Turkish.

In the treebank we use the label `compound:lvc` to link the nominal part of the predicate to the verb, and consider them as forming a single unit. This is similar to the approach taken in other languages in Universal Dependencies which have this feature.

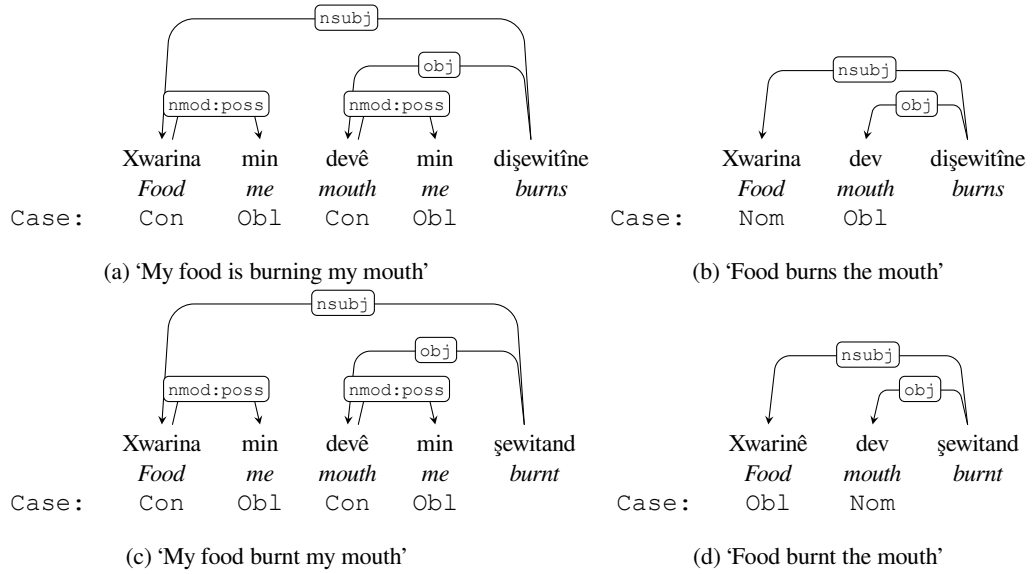
We use a number of diagnostics for determining if a given expression should be considered a light verb:

- “Is there another patient-like participant in the sentence aside from the nominal involved in the light verb construction?”
- “Is the nominal involved in the construction not inflected as if it were a simple argument to the verb? (i.e. is it inflected in the nominative case where it would otherwise be in the oblique?)”
- “Could this be considered a case of secondary predication?”
- “Are the constituents written together in the infinitive (e.g. in passive constructions, nominal use)”

An example is presented of a straightforward use of a light verb in Kurmanji. *Serokwezîr kuştina sivîlan şermezar dike.* “The Prime Minister condemns the killing of civilians.” The word *şermezar*, “shame”, is used together with the verb *kirin* to mean condemn, and the construction takes another argument as a direct object.



Unlike in some other languages, for example the Turkic languages, in Kurmanji these constructions may be discontinuous with an argument appearing between the verb and the nominal. For example: *Min bêriy**a te** kiriye*, ‘I missed you’ (lit. I did a before **of you**), has a construct case on the nominal part of



(a) 'My food is burning my mouth'

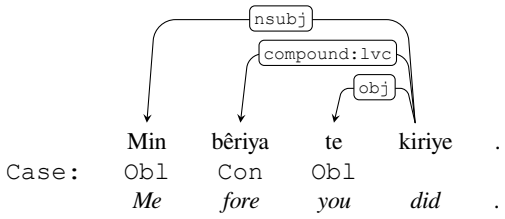
(b) 'Food burns the mouth'

(c) 'My food burnt my mouth'

(d) 'Food burnt the mouth'

Figure 2: Example of annotation of construct case. Note in (a) and (c) how the construct case overrides the verbal case government, which would have been nominative and oblique respectively (see §5.1).

the light verb construct, *bêriya* 'fore/before', which forms a noun phrase with the argument *te*, 'you'.



5.6 Future clitic

Future tense is expressed with present subjunctive inflection on the predicate and a *future clitic* after the subject. This clitic is usually in the form *dê*, which we consider to be its lemma in our annotation, but appears as *ê* after pronouns, and the pronoun and clitic often contract. For example compare examples (2a) and (2b).[10]

(2) a. *Hevalê zilam* ***dê*** *min*
Friend-CON man-OBL FUT me-OBL
bibîne.
see-FUT.3SG.

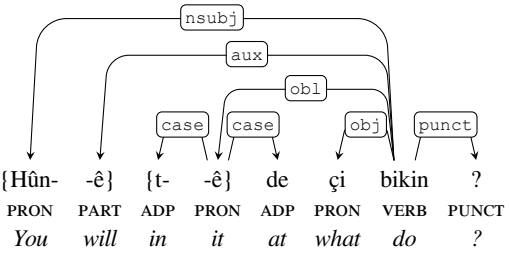
'The man's friend will see me.'

b. *Ezê biçim malê.*
I-FUT go-FUT.1SG home.

'I will go home.'

[10]The tag FUT stands for future tense, 3SG and 1SG stand for third and first person singular respectively.

The following example demonstrates the annotation of this feature for the sentence *Hûnê tê de çi bikin?* 'What will you do in there?'.



5.7 Pluperfect

The pluperfect tense is syntactically analytic but often contracts, e.g. *kirî bû* becomes *kiribû*. We currently represent this tense synthetically as this is how it is analysed by the morphological analyser. In the next version of UD Kurmanji we plan to split the tense up into its tokens of the main verb and the auxiliary *bûn*.

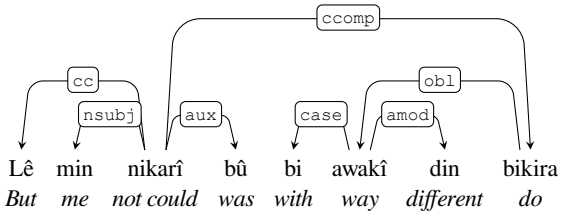
5.8 Subordination

Subordinate clauses are often formed with specific inflections, subjunctive in the present tense and what we have called 'optative' in the past.

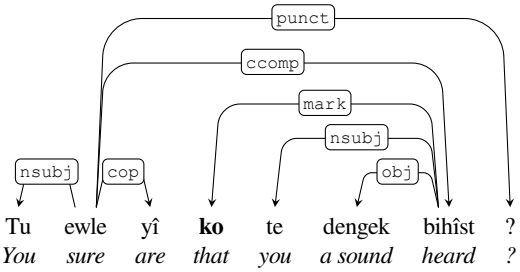
5.8.1 Complement clauses

In some cases subordination of finite clauses also occurs, with or without a complementiser. In the sentence, *Tu ji xwe ewle yî* ***ko*** *te dengekî fîkandinê û yê zencîrê bihîst?*, 'Are you sure **that** you heard a sound of whistling and a chain?' subordination is

done with the help of the complementiser *ku*, here written as *ko* as a result of dialect variation.



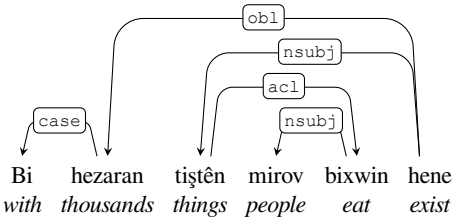
The verb form *bikira* in this sentence is an optative inflection of the verb *kirin*, 'to do'.



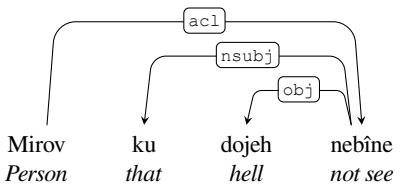
5.8.2 Relative clauses

Relative clauses can be introduced in three ways, which are not necessarily mutually exclusive.

Subjunctive mood: Here the mood of the subordinate clause indicates that the verb form is a nominal modifier. *Di xwezayê de bi hezaran tiştên mirov bixwin hene.* 'In nature things that people eat exist in thousands'.

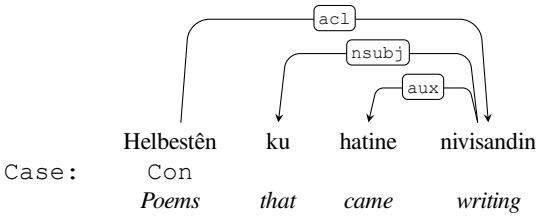


Relative pronoun: Very often a relative clause will be introduced with the use of a relative pronoun, usually *ku* 'that'/'who'. *Mirov ku dojeh nebîne*, 'a person who does not see hell'



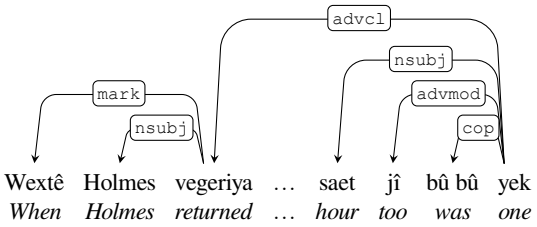
Note that like the English *that*, *ku* in Kurmanji is ambiguous between being a relative pronoun and a complementiser.

Construct case: A nominal in construct case is also a frequent way to introduce a relative clause. *Helbest**ê**n ku hatine nivisandin*, 'poems that have been written'.



5.8.3 Adverbial clauses

As in other Indo-European languages, in Kurmanji, adverbial clauses are usually introduced by subordinating or adverbial conjunctions. In the following sentence, *Wextê Holmes vegeriya...saet jî bû bû yek,* "By the time Holmes returned, the clock had struck one', the subordinationg conjunction *wextê* 'by' introduces the adverbial clause.



6 Parsing performance

In order to test the treebank in a real setting, we evaluated three widely-used popular dependency parsers: Maltparser (Nivre et al., 2007), UDPipe (Straka et al., 2016) and BiST (Kiperwasser and Goldberg, 2016). In addition we provide results for using the treebank for part-of-speech tagging using UDPipe, to be able to compare with Walther et al. (2010).

The BiST parser requires a separate development set for tuning. The set we used was the sample data from the shared task, this was 20 sentences, or 242 tokens. Both UDPipe and BiST parsers are also able to use word embeddings, we trained the embeddings using `word2vec` (Mikolov et al., 2013) on the raw text of the Kurdish Wikipedia. For Maltparser we used the default settings and for BiST parser we tested the MST algorithm.

We performed 10-fold cross-validation by randomising the order of sentences in the test portion of the corpus and splitting them into 10 equally-sized parts. In each iteration we held out one part for testing (75 sentences) and used the rest for training (675 sentences). We calculated the

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]**

Table 2: Preliminary parsing results for UDPipe and Maltparser. The numbers in brackets denote the upper and lower bounds found during cross-validation.

System	Lemma	POS	Morph
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]	93.0 [91.8, 93.8]	85.9 [84.2, 87.6]

Table 3: Performance of UDPipe for lemmatisation, part-of-speech and morphological analysis with the default parameters, and with an external full-form morphological lexicon.

labelled-attachment score (LAS) and unlabelled-attachment score (UAS) for each of the models using the CoNLL-2017 evaluation script.[11] The same cross-validation splits were used for training all three parsers.

The morphological analyser and part-of-speech tagger in UDPipe was tested both with and without an external morphological dictionary. In this case the morphological dictionary, shown in Table 2 as `[+dict]`, consisted of a full-form list generated from the morphological analyserdescribed in §4.2 numbering 343,090 entries.

The parsing results are found in Table 2. UDPipe is the best model, and adding the dictionary helps both POS tagging and parsing, an improvement of 2% LAS over the model without a dictionary.

For calculating the results for part-of-speech tagging, morphological analysis and lemmatisation, we used the same experiment but just looked at the results for columns 3, 4, and 6 of the CoNLL-U file. The results presented in Table 3 can be compared with the 85.7% reported by Walther et al. (2010) on 13 sentences. Predictably, in all cases adding the full-form list substantially improves performance.

7 Future work

The most obvious avenue for future work is to annotate more sentences. A treebank of 10,000 tokens is useful, and can be used for bootstrapping, but in order to be able to train a parser useful for parsing unseen sentences we would need to increase the number of tokens 6-10 fold.

We also think that there are prospects for working on other annotation projects based on the treebank, for example a co-reference corpus based on the short story.

There are a number of quirks in the conversion process from VISL to CoNLL-U, for example the language-independent longest-common-subsequence algorithm could be replaced with a Kurmanji specific one that would be able to successfully split tokens like *lê* into *l* and *ê*.

8 Concluding remarks

We have described the first syntactically-annotated corpus of Kurmanji Kurdish, indeed of any Kurdish language. The treebank was used as one of the *surprise language* test sets in the 2017 CoNLL on dependency parsing and is now released to the public. The corpus consists of a little over 10,000 tokens and is released under a free/open-source licence.

Acknowledgements

Work on the morphological analyser was funded through the 2016 Google Summer of Code programme and Prompsit Language Engineering with a contract from Translators without Borders.

We would like to thank Fazil Enis Kalyon, Daria Karam, Cumali Türkmenoğlu, Ferhat Melih Dal, Dilan Köneş, Selman Orhan and Sami Tan for providing native speaker insight and assisting with grammatical and lexical issues.

We would also like to thank Dan Zeman and Martin Popel for insightful discussions and the anonymous reviewers for their detailed and helpful comments.

References

Halil Aktuğ. 2013. *Gramera Kurdî – Kürtçe Gramer*. Avesta Publishing.

Purya Aliabadi, Mohammad Sina Ahmadi, Shahin Salavati, and Kyumars Sheykh Esmaili. 2014. Towards building kurdnet, the kurdish wordnet. In *Proceedings of the 7th Global WordNet Conference*.

Celadet Bedirxan and Roger Lescot. 1990. *Rêzimana Kurdî*.

Eckhard Bick and Tino Didriksen. 2015. Cg-3 – beyond classical constraint grammar. In *Proceedings of*

[11] `http://universaldependencies.org/conll17/evaluation.html`

the 20th Nordic Conference of Computational Linguistics, NODALIDA, pages 31–39. Linköping University Electronic Press, Linköpings universitet.

Kyumars Sheykh Esmaili and Shahin Salavati. 2013. Sorani Kurdish versus Kurmanji Kurdish: An Empirical Comparison. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 300–305.

M. L. Forcada, M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *TACL*, 4:313–327.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, and E. Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Chris Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Dan Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of Language Resources and Evaluation Conference (LREC'16)*.

Bişarê Segman. 1944. Dr. Rweylot. *Ronahî*, 24. Trad. Doyle, A. C. (1892) *The Adventure of the Speckled Band*.

Gary F. Simons and Charles D. Fennig, editors. 2017. *Ethnologue: Languages of the World*. SIL International.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Paris, France, May. European Language Resources Association (ELRA).

Wheeler M. Thackston. 2006. *Kurmanji Kurdish: A Reference Grammar with Selected Readings*. `http://www.fas.harvard.edu/~iranian/Kurmanji/index.html`.

Géraldine Walther, Benoît Sagot, and Karën Fort. 2010. Fast Development of Basic NLP Tools: Towards a Lexicon and a POS Tagger for Kurmanji Kurdish. In *International Conference on Lexis and Grammar*, September.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gökırmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria dePaiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. Conll 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada, August. Association for Computational Linguistics.

Appendix A. Format

Example sentence in VISL format, *Diviya bû tiştekî mihim qewimî biwa.* 'It must have been that something important had happened'

```
"<Diviya bû>"
        "divêtin" vblex plu p3 sg @root #1->0
"<tiştekî>"
        "tişt" n m sg con ind @nsubj #2->4
"<mihim>"
        "mihim" adj pst @amod #3->2
"<qewimî>"
        "qewimin" vblex iv pp @ccomp #4->1
"<biwa>"
        "bûn" vaux narr p3 sg @aux #5->4
"<.>"
        "." sent @punct #6->1
```