УДК 81’32

AN ASSESSMENT OF UNIVERSAL DEPENDENCY ANNOTATION GUIDELINES FOR TURKIC LANGUAGES

***Francis M. Tyers*[1], *Jonathan Washington*[2],**
***Çağrı Çöltekin*[3], *Aibek Makazhanov*[4]**
[1] School of Linguistics, Higher School of Economics, Moscow;
[2] Linguistics Department, Swarthmore College, Swarthmore;
[3] Seminar für Sprachwissenschaft, Universität Tübingen;
[4] National Laboratory Astana, Nazarbayev University, Astana
jonathan.washington@swarthmore.edu

Annotated corpora of three Turkic languages – Turkish, Kazakh, and Uyghur – were released as part of version 2 of the Free/Open-Source Universal Dependencies (UD) syntactic and morphological annotation guidelines. The objective of these guidelines is to provide consistent dependency annotation to facilitate cross-linguistic comparison.

This paper presents the current state of each of the three UD-annotated Turkic corpora, along with an evaluation of the performance of parsers trained on these corpora.

Overall, the UD annotation guidelines for Turkish, Kazakh, and Uyghur are fairly compatible – a testament to the careful design of the guidelines. However, the specific annotation guidelines for each of these languages were developed mostly independently; because of this, differences between the three standards exist. Moving forward with Turkic annotation standards in UD, attempts will be made to reconcile the differences. These differences are overviewed in this paper.

Furthermore, a number of issues in annotation have arisen and have yet to be resolved. Some of these issues require further investigation of the phenomena, and some require consultation within the UD community to determine whether solutions may be determined based on similar phenomena in other languages. A number of these open issues are discussed, including tokenisation (how to deal with words that include an orthographic space, or multiple words that do not include an orthographic space), the difference between core and oblique arguments of verbs, complex predicates (including structures where there is a combination of a non-finite form which governs argument structure and contributes to TAM and a finite-form which contributes to TAM and takes person agreement), multiple derivation (multiple causative or causative–passive combinations), and use of copulas instead of auxiliaries in what appear to be auxiliary constructions.

Keywords: Turkish; Kazakh; Uyghur; treebank; dependency grammar; Universal Dependencies.

ОЦЕНКА КРИТЕРИЕВ МОРФО-СИНТАКСИЧЕСКОЙ РАЗМЕТКИ ДЛЯ ТЮРКСКИХ ЯЗЫКОВ В ПРОЕКТЕ «UNIVERSAL DEPENDENCIES»

***Francis M. Tyers*[1], *Jonathan Washington*[2],**
***Çağrı Çöltekin*[3], *Aibek Makazhanov*[4]**
[1] Школа лингвистики, Высшая школа экономики, Москва;
[2] Департамент лингвистики, Суортмор-колледж, Суортмор;
[3] Школа лингвистики, Тюбингенский университет;
[4] Национальная Лаборатория Астана, Назарбаев Университет, Астана
jonathan.washington@swarthmore.edu

Аннотированные корпусы трех тюркских языков – турецкого, казахского и уйгурского – были выпущены в составе второй версии проекта «Universal Dependencies», предоставляющего свободно распространяемые рекомендаций к универсальной морфо-синтаксической разметке. Целью этих рекомендаций является предоставление единой схемы разметки для упрощения межязыкового анализа.

В настоящей работе описано текущее состояние каждого из трех тюркских корпусов размеченных по принципам UD, а также оценка эффективности синтаксических парсеров, обученных на этих корпусах.

Схемы разметки UD для турецкого, казахского и уйгурского языков во многом совместимы, что свидетельствует о тщательной проработке универсальности принципов UD. Однако конкретные рекомендации по аннотации для каждого из этих языков разрабатывались в основном независимо; из-за этого существуют различия между тремя стандартами. При дальнейшей разработке схем разметки для тюркских языков будут предприняты попытки сгладить данные различия, которые также рассмотрены в данной работе.

Кроме того, возник ряд вопросов по разметке определенных конструкций. Ответы на некоторые из этих вопросов требуют дальнейшего изучения природы соответствующих явлений в языке. В других случаях ответы могут быть получены на основе анализа схожих явлений в других языках; для этого потребуются консультаций с членами сообщества UD. В данной работе обсуждаются ключевые вопросы разметки, в частности: токенизация (считать ли слова, включающие в себя орфографические пробелы отдельными единицами разметки, и наоборот, разбивать ли несколько синтаксических слов, являющихся частью одного орфографического, на отдельные единицы разметки); разница между актантами и сирконстантами; сложные предикаты (включая структуры, где существует комбинация нефинитной формы, которая управляет аргументной структурой и несет временные и

аспектно-модальные функции, и финитной формы, которая также несет временные и аспектно-модальные функции и принимает личное окончание); множественная деривация (комбинации из нескольких каузативов и каузативов-пассивов); использование копулы вместо вспомогательного глагола в конструкциях, напоминающих сочетание главного и вспомогательного глаголов.

Ключевые слова: Турецкий язык; Казахский язык; Уйгурский язык; грамматика зависимостей; проект «Universal Dependencies».

1. Introduction

Universal Dependencies (UD, Nivre et al. 2016) is a Free/Open-Source set of guidelines for syntactic and morphological annotation of corpora, which aims to provide consistent dependency annotation to facilitate cross-linguistic comparison. In addition to the guidelines, annotated corpora are made available under a Free/Open-Source license.

This paper overviews recent work that has gone into making UD annotation guidelines for Turkic languages based on the UD standard. The current status of UD-annotated corpora of Turkic languages is overviewed in section 2. Three separate efforts have resulted in fairly compatible guidelines for Turkish (southwestern Turkic, §2.1), Kazakh (northwestern Turkic §2.2), and Uyghur (southeastern Turkic, §2.3), which is a testament to the careful design of the UD guidelines. However, because the specific annotation guidelines for each of these languages were developed mostly independently, differences between them exist, as described in section 2.5. Moving forward with Turkic annotation standards in UD, attempts will be made to reconcile the existing differences. In addition to efforts with these three languages and annotation standards, annotated corpora of some other Turkic languages have been begun as well (§2.4).

Furthermore, a number of issues in annotation have arisen and have yet to be resolved. Some of these issues require further investigation of the phenomena, and some require consultation within the UD community to determine whether solutions may be determined based on similar phenomena in other languages. A number of these open issues are discussed in section 4, including tokenisation (§4.1), the difference between core and oblique arguments of verbs (§4.2), complex predicates (§4.3), multiple derivation (§4.4), and use of copulas in auxiliary constructions (§4.5). Section 5 wraps up.

2. Current status

In this section we briefly describe treebanks of Turkic languages that have been or are about to be released in UD.

Table 1. **Turkic UD treebanks at a glance**

Treebank	Language	Sentences	Words	Annotation	Genre
Kazakh-UD	Kazakh	1 047	10 032	manual annotation	Wikipedia, fiction
IMST-UD	Turkish	4 660	48 093	semi-auto.	conversion news, social media
Turkish-PUD	Turkish	1 000	16 886	auto./manual	annotation translated news
Turkish-GK	Turkish	2 803	17 800	manual annotation	grammar examples
Uyghur-UD	Uyghur	100	1 662	semi-auto.	conversion fiction

In addition to the released Turkish (§2.1), Kazakh (§2.2), and Uyghur (§2.3) corpora, there has been some work on UD annotation of other Turkic languages (§2.4). Section 2.5 outlines the main differences between the annotation standards of the released corpora.

2.1 Turkish

Turkish is relatively well represented in the UD with two treebanks. The IMST-UD treebank (Sulubacak et al. 2016a) is the result of a semi-automatic conversion of the IMST treebank (Sulubacak et al. 2016b) which, in turn, was based on METU-Sabancı treebank (Oflazer et al. 2003). The second Turkish treebank, Turkish-PUD, in the official UD repository is part of the parallel treebanks released during CoNLL 2017 UD parsing shared task (Zeman et al. 2017). Besides these two treebanks, the treebank reported in Çöltekin 2015 (Turkish-GK) is annotated for the purpose setting UD annotation guidelines for Turkish. To cover a wide range of morphosyntactic phenomena, the Turkish-GK treebank annotates the example sentences from comprehensive grammar book. This treebank follows UD version 1.3 annotation scheme, and currently not converted to version 2.0.

2.2 Kazakh

Kazakh is represented in UD by a single treebank (Makazhanov et al. 2015; Tyers and Washington 2015), which was first released in UD v1.3, and at the moment of writing contains 1109 trees (sentences) and a total of 10894 tokens. The annotation scheme of the treebank defines 16 UD POS tags, 45 "category=value" feature pairs, and 34 dependency relations of which four are language-specific. Tokenisation and morphological processing strategies in the Kazakh UD treebank follow the principles of Turkic lexica as defined by the Apertiun project[1]. One reason for this is to keep the UD corpus compatible with the morphological analysers developed by the Apertium Turkic working group.

Currently the treebank is partially compatible with UD v2.0 standard, with the choice of head direction in some constructions being one of the major discrepancies. The standard requires coordination and some compounds (e.g. names) to be left-headed, while the treebank developers believe that in Kazakh (and other Turkic languages) such constructions should be right-headed due to the placement of morphological locus, which is exclusive to the last (rightmost) element of such constructions. So far this issue has been resolved by an intermediate conversion step, where initially the annotation is performed in a right-headed fashion, and at the time of release a special script flips the heads of the constructions in question.

2.3 Uyghur

In Aili et al. (2016b), a treebank for Uyghur with 20,000 tokens is described. Tokens fit into one of 12 part-of speech categories and there are 137 morphological tags. There are 23 total dependency relations, with adjuncts classified by morphological case. In co-ordination, the conjunction is attached to the following conjunct and the preceding conjunct is attached to the following one (so-called 'head-final' conjunction).

Aili et al. (2016a) present a conversion of the Uyghur dependency treebank Aili et al. (2016b) to Universal Dependencies. They used some default mapping rules to convert the parts of speech and de-

[1] http://wiki.apertium.org/wiki/Turkic_lexicon

pendency relations, and then some limited rules based on the part of speech of the head to distinguish between ambiguous relations (for example mapping `att → {amod, det, nummod}`. The treebank contains surface forms, parts of speech and dependency relations, but no lemmas or morphological features.

2.4 Other

Ageeva and Tyers (2016) present two small treebanks for Tuvan and for Crimean Tatar of approximately 1,000 tokens each for use in testing a method of cross-lingual dependency parsing. They show that it is possible to take advantage of a morphological analyser and a treebank for another language in order to learn an improved delexicalised parser.

2.5 Main differences

At present, there are a number of differences in the dependency annotation standards for Kazakh, Turkish, and Uyghur. Quite a few of these differences are in the morphological annotation (part-of-speech tags and morphological features), but there are a handful of differences in tokenisation (how to approach words that include an orthographic space, or multiple words that do not include an orthographic space) and dependency annotation as well. In general, the Kazakh and Turkish annotation standards are more compatible with one another than either is with Uyghur.

One example of a difference in part-of-speech tagging is how locational pronoun-derived adverbials are represented. In Turkish, words like *nerede* 'where' and *nereden* 'from where' are labelled as `PRON` (with the appropriate case indicated in the morphology features), and hence usually have the dependency relation of nominal adverbials, `obl`. In Kazakh, the corresponding words *қайда* 'where' and *қайда* 'from where' are labelled as flat adverbs, or `ADV`, and hence have dependency relations of `advmod`. Which analysis is more appropriate is not clear: the fact that they are pronouns with case suffixes in both languages argues for their annotation as pronouns, while the fact that these pronouns are defective (they can't take all case suffixes) in each language argues for their analysis as grammaticalised adverbs.

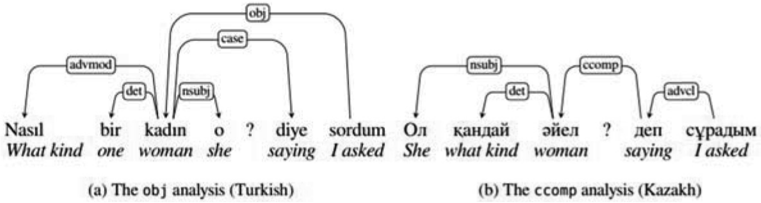


Figure 1. Two alternative analyses for speech with the adverbial "quotative word" use of the verb de- 'say'. Both sentences mean "I asked, 'what kind of woman is she?'" Analysis 1a shows the de- verb form as a case marker to the clause it governs, which is in turn an object of the main verb. In the analysis in 1b, the speech verb is treated as a verbal adverb adjunct to the main verb, with its own clausal complement

The current Uyghur corpus does not have annotation for morphological features, and there are some notable differences between Kazakh and Turkish standards. One of these is the annotation of `Person=3` in Turkish for any nominal–since as they trigger third-person agreement morphology as subjects and possessors. Kazakh does not have this feature. Similarly, Turkish annotates for `Polarity` values (e.g., on verb forms) of both `Pos` 'positive' and `Neg` 'negative', whereas Kazakh only indicates polarity if the value is negative.

Table 2. **Language specific relations: the tick mark (✓) means that the relation is found in the treebank; the – mark means that the relation could apply, but is not applied at present; and the asterisk (*) means that we consider this relation to be an erroneous classification**

Relation	Comments	Kazakh	Turkish	Uyghur
acl:poss	Adnominal modification with possessive	✓	–	–
acl:relcl	Adnominal modification with verbal adjective	✓	–	–
advmod:emph	Adverbial emphasiser (mostly -dA)	–	✓	✓
aux:q	Question word, -mI	–	✓	–
compound:lvc	Light verb	✓	✓	✓
compound:redup	Reduplication compound	–	✓	✓
flat:name	Proper name	✓	–	–

iobj:caus	Causee	✓	–	–
nmod:abl	Oblique in the ablative	*	*	✓
nmod:cau	Causee	*	*	✓
nmod:clas	Noun-noun compound	*	*	✓
nmod:comp	Nominal modifier [mostly ablative]	–	–	✓
nmod:poss	Genitive possessive modifier	✓	✓	✓
nmod:tmod	Time modifier	–	–	✓
obl:own	Owner in -DA	✓	–	–

In terms of tokenisation, the Turkish tokenisation standard always considers denominal adjectives formed with -/lI/ to be a noun followed by an adposition (i.e., two tokens), while in Uyghur these words are treated as lexicalised adjectives (i.e., as one token). The Kazakh treebank varies between these two approaches, currently in a somewhat unprincipled way. Another difference in terms of tokenisation is treatment of the so-called -/ki/ affix, two uses of which are the formation of the attributive locative (-/DAki/ in Turkish, -/DAGI/ in Kazakh) from the locative case form (-/DA/ in both) and the formation of the substative genitive (- /(n)Inki/ in Turkish and -/Niкi/ in Kazakh) from the genitive suffix (-/(n)In/ in Turkish and -/NIң/ in Kazakh). In all three languages, forms with these "compound" affixes are annotated with the appropriate relation (`amod` for attributive locative), but in Kazakh and Uyghur, these forms are not analysed as containing a separate -/ki/ suffix. That is, in Kazakh and Uyghur, there is a single token, while in Turkish, a separate -/ki/ token has a case dependency to the noun.

One difference in annotation of dependency relations is the treatment of the adverbial "quotative word" (Turkish *diye*, Kazakh *деп*, Uyghur دەپ). In all three languages, this form, morphologically speaking, is a verbal adverb (participle) form of the verb *de-* 'say', though in modern Turkish, the -/(y)A/ participle used in *diye* is not productively used as a verbal adverb. In Kazakh and Uyghur treebanks, it is analysed this way–thatis, it receives an `advcl` analysis (dependent on the main clause it comes in) and is labelled a `VERB`, with its relation to the head of its clausal complement labelled as a `ccomp`. This is shown in figure 1b. In the Turkish treebank, however, the speech verb is treated as an `ADP` and has a case dependency relationship to the head of the “quoted” phrase, as shown in figure 1a.

There are also a number of differences with how language-specific relations are used. Table 2 presents a summary of the language-specific relations used in each treebank, and the applicability to the other treebanks.

3. Parsing performance

All three treebanks were included in the 2017 CoNLL shared task on Universal Dependency parsing from raw text data (Zeman et al. 2017). The results for the Turkic languages are presented in Table 3. LAS and UAS stand for labelled-attachment score and unlabelled-attachment score respectively, while CLAS stands for content-labelled attachment score (Nivre and Fang 2017).

Table 3. **Parsing performance in the CoNLL shared task. The column Train indicates the number of tokens in the training data, and the column Dev indicates the number of tokens in the development data. Note that there were no separate training sets for Turkish and Turkish-PUD; the latter is only used as a test set for parsers trained on Turkish training data**

Language	Train	Dev	Winning team (LAS)	UAS	LAS	CLAS
Kazakh	0	529	Dumitrescu et al. (2017)	45.72	29.22	25.14
Turkish	38 082	10 011	Dozat et al. (2017)	69.62	62.79	60.01
Turkish-PUD	38 082	10 011	Björkelund et al. (2017)	59.35	38.22	32.32
Uyghur	0	1662	Björkelund et al. (2017)	60.57	43.51	34.07

It is interesting to note the difference between the parsing performance on the Turkish section of the parallel treebank (Turkish-PUD) and on the IMST treebank (the main UD treebank for Turkish). Both of these treebanks have been converted from other treebank formalisms: the METU-Sabancı formalism in the case of the IMST treebank and Google Universal Dependencies in the case of the Turkish section of the parallel treebank.

It is also curious as to why the Kazakh and Uyghur numbers are so different despite the data size being similar (that is, while the Uyghur dev set had three times as many tokens, it didn’t have lemmas or any morphology annotation). One explanation could be that there was a lot more data sparsity when tagging Kazakh as opposed to Uyghur. It's also striking that Björkelund et al. (2017) got slightly better results for Uyghur than Turkish-PUD, despite a much smaller development corpus size. It should be mentioned that in addition to the training and development data, the shared task included at least 10 000 tokens of both Kazakh and Uyghur testing data.

4. Open questions

In this section we discuss certain phenomena in Turkic languages that present challenges during annotation. Some of those phenomena, e.g. multiple derivation, can be fairly well understood, but are difficult to handle adequately given the present UD annotation guidelines. The nature of others may require further investigation within the dependency grammar formalism.

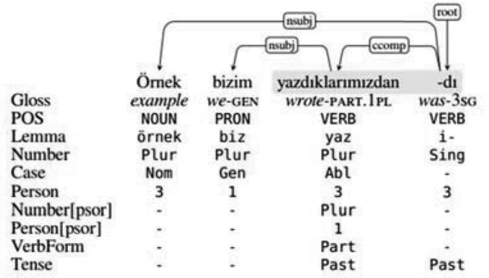
4.1 Tokenisation

One of the guiding principles of Universal Dependencies is about its *lexicalism* (Nivre et al. 2016). That is, *words* are the basic units of annotation. The guidelines explicitly state that *word* here refers to *syntactic words*. It is allowed for orthographic words to be split when it is necessary for the syntactic analysis.

In earlier Turkish dependency parsing/annotation work, on the other hand, the words are split at all *derivation boundaries*, introducing a syntactic word (often called an *inflectional group* in Turkish NLP literature) for each derivational morpheme. For example, the Turkish word *sınırlandırılabilecek* 'that can be limited' can be represented as six syntactic words, delimited by ˆDB 'derivation boundary', as shown in (1).

```
(1) sınır+Noun+A3sg+Pnon+Nom
    ˆDB+Verb+Acquire
    ˆDB+Verb+Caus
    ˆDB+Verb+Pass
    ˆDB+Verb+Able+Pos
    ˆDB+Adj+AFuttPart
```

Although derivation in Turkic languages can be quite productive, as exemplified by (1) above, arguing for necessity of this level of word segmentation is not always practical. Present Turkic UD treebanks segment the words when parts of the word may have conflicting morphological features and/or parts of the word can participate in different/conflicting syntactic relations (Çöltekin 2016). In (1) above, none of the syntactic words are necessary since UD morphological features can mark the effect of each morpheme and none of the parts can participate in different syntactic relations–i.e., the parts cannot be modified by other words or ambiguously head other words in a sentence. However, there are also examples where the split is necessary. For example, failing to split the copular suffix in Figure 2 results in two nsubj relations headed by the same word[1]. Furthermore, the same word would be assigned both Number=Plur and Number=Sing features.



	Örnek	bizim	yazdıklarımızdan	-dı
Gloss	*example*	*we*-GEN	*wrote*-PART.1PL	*was*-3SG
POS	NOUN	PRON	VERB	VERB
Lemma	örnek	biz	yaz	i-
Number	Plur	Plur	Plur	Sing
Case	Nom	Gen	Abl	-
Person	3	1	3	3
Number[psor]	-	-	Plur	-
Person[psor]	-	-	1	-
VerbForm	-	-	Part	-
Tense	-	-	Past	Past

Figure 2. Dependency analysis of the sentence *Örnek bizim yazdıklarımızdandı* 'The example was from the ones we wrote'. Note how on the verbal form *yazdıklarımızdandı* there would be two values for Person, Number and Tense were the copula not split from the non-finite form

How to treat very productive derivational suffixes, which attach to phrases rather than single words, is also a challenge. These include the suffixes -/LI/ 'with', -/sIz/ 'without', and -/LIK/ '-ness, -ed', which appear in most Turkic languages. Very many forms that include these suffixes are lexicalised, for example *evsiz* 'homeless' (lit. 'without house'), *evli* 'married' (lit. 'with house'), *gözlük* 'spectacles' (lit. 'eye-

[1] Note that an analysis of this sentence more in line with the annotation standards for Kazakh would have yazdıklarımızdan as the root, with dı as a cop dependent of it, but this again results in the problem of having two nsubj relations headed by the same word. Issues like this are recognised by v2 of the UD standard (cf., http://universaldependencies.org/v2/copula.html) and affect non-Turkic languages as well.

ness')[1], but in some cases, for example *бір палаталы* 'one chambered, unicameral', it would be advantageous to consider the suffix separately since *бір* 'one' modifies the stem, not the whole form. There are potentially ambiguous examples as well, such as *iki gözlük*, which can technically mean either 'two glasses' or 'for two eyes' (e.g., the value of something–though this usage would be rare)[2], depending on what level the *lük* suffix is interpreted at. One possible solution would be to only split if the word to which the suffix is attached has its own modifiers of a certain type, although this sort of structural difference is difficult to segment in an NLP pipeline.

Another associated issue is related to syntactic words which contain multiple surface words. An example is the Turkish question marker which, when attached to predicates, may also carry some of the morphological features of the predicate. UD currently does not explicitly support syntactic words spanning multiple tokens, though the Kazakh treebank implements some things this way. Some rudimentary solutions exist in the present UD scheme–e.g., the `goeswith` relation or considering a space-separated token a single token–but ad hoc use of these without a standard could cause inconsistencies between Turkic language treebanks. Some issues related to syntactic words containing multiple surface words are discussed further in section 4.3.

Although some general guidelines exist for segmentation of words, there is a need for widely accepted, more concrete rules to ensure consistency among Turkic languages, and even among treebanks of the same language.

4.2 Core and oblique

In the UD v2 standard, the dependency relations `obj`, `iobj`, and `obl` are differentiated in the following way: `obj` is the most core element of a verb that is not its subject (i.e., a direct object), `iobj` is the

[1] It should be mentioned that it's not clear that these examples aren't understood by native speakers of Turkish as compositional and productively formed. Instead, perhaps this interpretation relies on the translation of these words to other languages (a poor criterion!) – it is not necessarily "metaphorical" (at least historically) that evli should mean 'married'.

[2] A reading that might be found in a wider range of real sources might be 'two-division' or 'two-room', based on another meaning of göz. In any case, any interpretation of such forms will depend on the context and whether an established lexicalised meaning exists.

next most core element that isn't a subject or direct object, and oblique is a non-core object. This relies on the notion of a difference between core and non-core elements, or complements versus adjuncts, respectively.

In Turkic languages, there does not seem to be a simple and clear way to delineate complements and adjuncts. No element of a verb phrase is absolutely required to be included in a grammatical utterance, not even the subject. While agreement marking will show the existence of a semantically present subject, even if not included in the sentence, Turkic languages do not mark object or indirect object agreement on the verb. Furthermore, since most of the cases have a very wide range of uses, many phrases can be used in any verb phrase, although with a different interpretation depending on the verb.

It seems clear, at least, that typical "accusative direct objects" (and morphologically unmarked indefinite direct objects) should be annotated with the `obj` relation. However, there is currently only one test that we can use to justify this and other relations: if the element participates in case promotion or demotion when the verb is made passive or causative, we consider it a core argument, to be labelled with `obj` if it seems "more core" and `iobj` if there is another element labelled obj. If the case marking does not change when the verb is made passive or causative, then the element is considered oblique, and receives the `obl` dependency relation. A more apt solution may exist, but has yet to be identified.

4.3 Complex predicates

In this subsection we discuss verbal (i.e., *non-copular*) complex predicates. Such predicates consist of two or more orthographic words that together convey single meaning, which is different from meanings (if any) of those words taken separately. Sometimes it is not at all clear how to classify the relationship between the constituents of complex predicates. For instance, a common Kazakh expression *пайда бол*, meaning appear or be established consists of what appears to be a noun *пайда* ('benefit', when used on its own as a noun) and a verb *бол* ('be' or 'finish'), which in this particular case loses its habitual copular and auxiliary functions. Thus, a verb that normally takes no arguments[1] in

[1] In UD copulas are subordinated to nominal predicates for the sake of cross-linguistic consistency (http://universaldependencies.org/u/overview/syntax.html).

this case governs what appears to be a noun; the question is with what syntactic relation.

Depending on the nature of their constituents, complex predicates in question can be roughly classified into three categories: (i) non-verbal + verbal; (ii) verbal + non-verbal; (iii) verbal + verbal. Assuming that predicates are finite (i.e., non-clausal), in all of these cases the rightmost constituent carries a personal agreement marker (sometimes covert), and in the latter two categories the first verbal constituent is usually non-finite[1] and contributes to TAM. Also, in all of the cases 'particles' and conjunctions may be inserted between the constituents at will – e.g., compare *пайда болды* 'it appeared' and *пайда да болды* 'and it also appeared'.

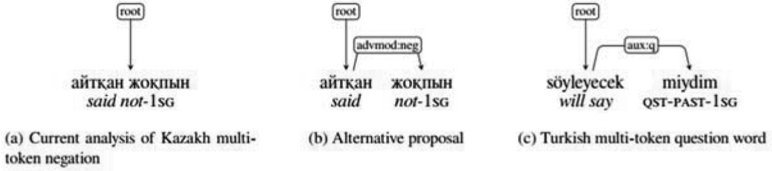
The first category of complex predicates (non-verbal + verbal) in certain UD treebanks (including Kazakh) is sometimes handled as a special sub-type of compounds, namely a light verb construction. This solution, while possible, relies on meaning, which is undesirable. There are two alternatives: (i) treat such constructions as a single space-separated token (which in some cases is done in the Apertium Kazakh lexicon); (ii) sacrifice the meaning and treat such constructions just as normal verb-argument or nominal-copula relations. Both alternatives have pros and cons. While the first one could be accommodated at the level of morphological analysis and tagging, it is not clear how to handle embedded 'particles' and conjunctions. As for the second alternative it just seems wrong to impose literal (usually absurd) meaning on otherwise meaningful constructions[2].

The second category of complex predicates (verbal + non-verbal) corresponds in Kazakh to negation of finite verbs which appears to consist of two tokens. In this construction, the first element (before the space) is morphologically a verbal noun or adjective ending in -/GАн/ and the last element is either *жоқ* 'non-existent' or *емес* 'not' – e.g., *айтқан жоқпын* 'I did not say'. Currently in the Kazakh treebank this case is handled at the morphological analysis step, and the construc-

[1] The only exception that we are aware of is the non-morphological negation, where (at least in Kazakh) the initial verbal constituent may agree with the subject, e.g. мен олай айтқаным жоқ vs мен олай айтқан жоқпын.

[2] Especially if a non-verbal constituent has no lexical meaning on its own and exists only in this sort of expression – e.g., міз бақпа 'pay no attention' or мiсе тұт 'be satisfied', where міз and мiсе do not exist as lexical items outside of these constructions.

tion is treated as a single token, as shown in figure 3a. It is currently unclear what to do when function words are embedded between the elements of constructions like this. Alternatively this sort of construction could be treated as analytic negation with *жоқ* or *емес* being considered negation words and subordinated to the leading verb with the relation `advmod:neg`, as shown in figure 3b. This approach however leads to non-verbal entities carrying personal agreement markers, which is undesirable. Although, this happens in the current conversion of the Turkish treebank with the question word -/mI/ which can carry person/number agreement and tense, as demonstrated in figure 3c.



(a) Current analysis of Kazakh multi-token negation

(b) Alternative proposal

(c) Turkish multi-token question word

Figure 3. Some examples of Kazakh multi-token negation (‘I didn’t say’) and the current analysis of Turkish multi-token question word (‘Would I say?’)

The third category of complex predicates (verbal + verbal) includes constructions which appear similar to auxiliary verb constructions, but which are probably best thought of as verbal adverb adjunct of main verb. The trailing finite verb does not contribute to TAM (tense, aspect, and mood – which typical auxiliaries in Kazakh convey), and the meanings of these combinations “feel” lexicalised (though it’s unclear whether this is just due to how they translate to other languages). Some examples include *болып табылады* ‘is found to be’, *болып саналады* ‘is considered’, *атап өтті* ‘mentioned’, *алып келді* ‘brought’, etc. In such constructions the preceding verbs assume a form of -/(I)п/ verbal adverb which can be followed by an auxiliary. The trailing verbs, however, are not always in the closed class of auxiliaries – and the ones that can be auxiliaries do not convey the normally associated auxiliary meaning, such as contributing to TAM. Both verbal elements give a combined meaning to the entire construction, e.g. *ата* ‘name (V)’ + *өт* ‘pass (V)’ combine as *атап өт* ‘mention’. Currently some of these constructions are treated as a single token in the Kazakh treebank due to the fact that they are lexicalised in the morphological analyser used to preprocess text for the treebank, but is designed for use in machine translation pairs where such lexicalisation is useful. Because of this single-token analysis, the previously mentioned problem of function word embedding exists for these forms. However, other occurrences in the treebank are treated as a main verb with an advcl dependent. One disadvantage of this is that it does not match intuitions about the lexical semantics of these constructions, although, again, perhaps these intuitions are based on the translation of these forms to other languages. Another disadvantage of treating the verbs separately is that in some cases it can result in crossing dependencies, as shown in figure 4.

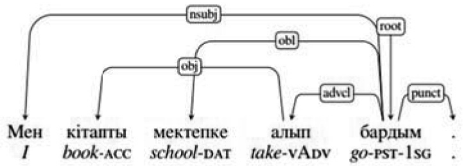


Figure 4. A separate-word analysis of a V+V compound verb in Kazakh, showing overlapping dependencies. The sentence translates to "I brought the book to school." Here, the verb *бар* 'go' has the oblique dependent *мектепке* 'to the school', while the verb *ал* has the direct object *кітапты* 'the book' (accusative). Note that a different order of the words, *Мен кітапты алып мектепке бардым* would have a slightly different meaning, 'I took the book and went to school' or 'Taking the book, I went to school'–and may not entail that the book ended up being brought to school as the depicted sentence does

The following facts all point to the interpretation of these verbs operating as a single, compound unit: that both verbs can contribute to the argument structure of the entire phrase, that the semantics are not always entirely compositional (but each verb usually contributes something), that the phrase seems to represent one event and not two, and that the verbs together share a single TAM reading. One possibility for how to deal with this would be to annotate these constructions with compound (or a subtype of the compound relation, e.g. compound:v), with the finite verb governing the non-finite one.

Another case that could be considered to fall under this category of complex predicates has not actually occurred in the Kazakh treebank, but is fairly frequent in speech: when a Russian infinitive is followed by the verb *ет* 'do, make'– e.g. *звонить ет* 'make a telephone call', *обжаловать ет* 'appeal to a higher court', etc. Due to the introduction of a foreign word, these constructions could potentially be handled with a special dep relation that preserves structure but bears no particular grammatical function. Thus, the Russian infinitive could be tagged as X (uncategorised word) and be subordinated to the trailing verb with the relation dep. If a function word is embedded in between, it can be subordinated to the Russian infinitive with the same relation.

4.4 Multiple derivation

Many linguistic phenomena that are commonly expressed by syntactic means – e.g., relations between words – are expressed by morphological features in Turkic languages. In particular, Turkic verbs can be inflected for a range of affixes expressing features like tense, aspect, modality, voice and subject agreement. The UD specifications allow representing most of these features, and the changes in version 2.0 improved this representation considerably. However, in some cases the UD morphology specification is still sub-optimal for expressing some morphological phenomena.

The issue mainly arises when multiple Aspect, Mood and Voice features are present on the same verb. The Turkish examples in (2) include multiple Voice (2a) and Aspect (2b) features on a verb.

```
(2) a. bekle -t   -il  -iyor
       wait  CAUS PASS PROG
       'being stalled (=caused to wait)'
    b. oku  -yuver -iyor
       read RAPID  PROG
       'he/she is reading quickly'
```

In (2a), the verb has both passive voice and causative voice. Similarly in (2b), the affixes indicate that the action is done quickly, and it is in progress, both of which are typically defined as *aspect* (Göksel and Kerslake 2005). While the UD version 2.0 specifications allow marking each of these feature values individually, there is no clear way to mark multiple values for a single feature. In Turkish UD treebank, these words are marked using language-specific feature values. For example, the morphological annotation of (2a) includes Voice=CauPass, and the multiple aspect suffixes in (2b) are indicated by Aspect=ProgRapid.

A related issue is repetition of some of these features. Notably, the Turkish causative marker can be attached to a verb multiple times without a principled limit, indicating a chain of causation[1]. Similarly, Turkish possibility/ability mood marker can be repeated two times in certain contexts. It is worth noting that this may also arise in non-Turkic languages – see e.g., Ainu in Senuma and Aizawa (2017).

Note that although the above method encodes the relevant information, it makes it difficult for an automated system (e.g., a parser), since the symbol `CauPass` does not clearly indicate the features `Cau` or Pass unless special attention is paid for this non-standard notation. Since some of these combinations are rare[2], it is difficult for a machine learning method to automatically discover that `CauPass` is equivalent to having both features marked individually. Similarly, a researcher, for example, looking for causative verbs in the language using a treebank search tool will likely to be misled by this ad hoc representation.

4.4 Use of copulas with non-finite verb forms

One issue that has arisen recently is how to analyse the use of copulas (as opposed to auxiliary verbs) with non-finite verb forms that occur together with auxiliary verbs. There appear to be cases of this in many Turkic languages. Normal non-finite form + auxiliary forms are straightforwardly dealt with in UD, as in figure 5.

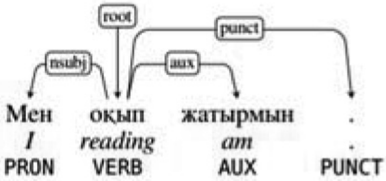


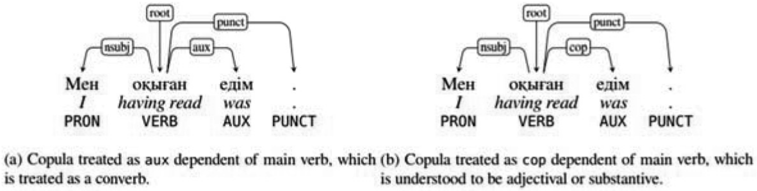
Figure 5. Dependency analysis of an auxiliary construction in Kazakh

In both Kazakh and Turkish, a number of finite verb forms are formed with what appears to be a verbal noun or verbal adjective form,

[1] Although real-life use is often limited, and multiple causative markers often (but not always) indicate emphasis rather than multiple levels of causation. In the Turkish-UD treebank there are no examples of multiple causative, but Turkish-GK includes examples with two causative suffixes, also in combination with the passive suffix.

[2] Of 9113 verbs in Turkish-UD treebank, Voice=CauPass is the most common multiple-feature marking with 115 occurrences. Others include 5 instances of Mood=CndPot, 4 instances of Mood=GenNec and Mood=DesPot, 2 instances of Aspet=DurPerf, and single instances of Aspect=ProgRapid and Mood=NecPot.

followed by normal copula agreement[1]. This includes forms like Turkish *okumuşum* ‘I read (past)’ (with perfect form *okumuştum* ‘I had read’) and *okurum* ‘I read (non-past)’ (with perfect form *okurdum* ‘I would read’) and Kazakh *оқығанмын* ‘I read (past)’ (with perfect form *оқыған едім* ‘I had read’) and *оқырмын* ‘I may read’ (with perfect form *оқыр едім* ‘I would read’). These structures are entirely parallel, although the Kazakh forms have a space between the verb form and the past form of the copula. This construction lends itself to a number of different analyses, as shown in figure 6[2]. There are also, in a smaller set of Turkic languages, “auxiliary” constructions that appear to be composed of a copula along with a form that cannot ever be verbal nouns or verbal adjectives and can only operate as a non-finite form together with an auxiliary. Because these look more like true auxiliary phrases, it’s clearer how they might be treated. One analysis is shown in figure 7.



(a) Copula treated as aux dependent of main verb, which is treated as a converb.

(b) Copula treated as cop dependent of main verb, which is understood to be adjectival or substantive.

Figure 6. Two different analyses of an apparent auxiliary construction in Kazakh that consists of a verbal noun or adjective and a copula, glossed as ‘I had read’. These analyses differ in how they view the copula auxiliary: either as the auxiliary in an auxiliary verb construction or as the copula in a normal copula predicate which happens to have a substantive or attributive verb form as the predicate

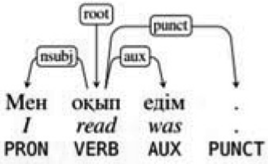


Figure 7. One analysis of an apparent auxiliary construction in Kazakh that consists of a non-finite form and a copula, glossed roughly as ‘I had read’

[1] Note that this is distinct from the obvious verbal nouns in forms like Turkish okumaktayım or Kazakh оқудамын – both meaning “I am reading” (with a long-term or habitual sense), and can be analysed as a verbal noun followed by a locative suffix, followed by a copula with person agreement.

[2] Note that the part of speech of copulas is always AUX in UD.

The set of choices around copula-as-auxiliary constructions includes several options for both dependency relations and morphological annotation, and Turkic languages have different orthographic strategies regarding tokenisation. This issue would be especially good to solve before further annotation.

5. Concluding remarks

We have presented an overview of the current status of the Turkic languages within the Universal Dependencies project and drawn attention to a number of inconsistencies that remain. We hope that this serves to inform and direct future research in the area of Turkic dependency parsing and Turkic language technology in general. After careful analysis we are convinced that the majority of substantial differences in the annotation schemes are a result of conversion from different grammatical traditions as opposed to real significant grammatical differences.

Acknowledgements

The authors would like to thank the UD community for thoughtful discussion and input on a range of issues discussed in this paper. We would also like to thank Deniz Uysal and Tolgonay Kubatova for help with native speaker judgements. The work of Aibek Makazhanov was supported by Nazarbayev University under the research grant №129-2017/022-2017.

REFERENCES

1. Ageeva, Ekaterina and Francis M. Tyers (2016). “Combined morphological and syntactic disambiguation for cross-lingual dependency parsing”. In: Proceedings of TurkLang 2016.

2. Aili, Mairehaba, Weinila Mushajiang, Tuergen Yibulayin, A. Kahaerjiang, and Yan Liu (2016a). “Universal dependencies for Uyghur”. In: Proceedings of WLSI/OIAF4HLT. Osaka, Japan, pp. 44–50.

3. Aili, Mairehaba, Aziguli Xialifu, Maihefureti, and Saimaiti Maimaitimin (2016b). “Building Uyghur Dependency Treebank: Design Principles, Annotation Schema and Tools”. In: International Workshop on Worldwide Language Service Infrastructure, pp. 124–136.

4. Björkelund, Anders, Agnieszka Falenska, Xiang Yu, and Jonas Kuhn (2017). “IMS at the CoNLL 2017 UD Shared Task: CRFs and Perceptrons Meet Neural Networks”. In: Proceedings of the CoNLL 2017 Shared Task:

Multilingual Parsing from Raw Text to Universal Dependencies. Vancouver, Canada: Association for Computational Linguistics, pp. 40–51.

5. Çöltekin, Çağrı (2015). “A grammar-book treebank of Turkish”. In: Proceedings of the 14th workshop on Treebanks and Linguistic Theories (TLT 14). Ed. by Markus Dickinson, Erhard Hinrichs, Agnieszka Patejuk, and Adam Przepiórkowski. Warsaw, Poland, pp. 35–49.

6. Çöltekin, Çağrı (2016). “(When) do we need inflectional groups?” In: Proceedings of The First International Conference on Turkic Computational Linguistics. Konya, Turkey.

7. Dozat, Timothy, Peng Qi, and Christopher D. Manning (2017). “Stanford’s Graph-based Neural Dependency Parser at the CoNLL 2017 Shared Task”. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Vancouver, Canada: Association for Computational Linguistics, pp. 20–30.

8. Dumitrescu, Stefan Daniel, Tiberiu Boroş, and Dan Tufiş (2017). “RACAI’s Natural Language Processing pipeline for Universal Dependencies”. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Vancouver, Canada: Association for Computational Linguistics, pp. 174–181.

9. Göksel, Aslı and Celia Kerslake (2005). Turkish: A Comprehensive Grammar. London: Routledge.

10. Makazhanov, Aibek, Aitolkyn Sultangazina, Olzhas Makhambetov, and Zhandos Yessenbayev (2015). “Syntactic Annotation of Kazakh: Following the Universal Dependencies Guidelines. A report”. In: Proceedings of the 3rd International Conference on Turkic Languages Processing, pp. 338–350.

11. Nivre, Joakim and Chiao-Ting Fang (2017). “Universal Dependency Evaluation”. In: Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), pp. 86–95.

12. Nivre, Joakim, 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).

13. Oflazer, Kemal, Bilge Say, Dilek Zeynep Hakkani-Tür, and Gökhan Tür (2003). “Building a Turkish treebank”. In: Treebanks: Building and Using Parsed Corpora. Ed. by Anne Abeillé. Springer. Chap. 15, pp. 261–277.

14. Senuma, Hajime and Akiko Aizawa (2017). “Toward Universal Dependencies for Ainu”. In: Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017). Gothenburg, Sweden, pp. 133–139.

15. Sulubacak, Umut, Memduh Gökırmak, Francis Tyers, Çağrı Çöltekin, Joakim Nivre, and Gülşen Eryiğit (2016a). “Universal Dependencies for Turkish”. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan, pp. 3444–3454.

16. Sulubacak, Umut, Tuğba Pamay, and Gülşen Eryiğit (2016b). “IMST: A revisited Turkish dependency treebank”. In: Proceedings of the 1st International Conference on Turkic Computational Linguistics (TurCLing). Konya, Turkey.

17. Tyers, Francis Morton and Jonathan North Washington (2015). “Towards a free/open-source dependency treebank for Kazakh”. In: Proceedings of the 3rd International Conference on Turkic Languages Processing, pp. 276–289.

18. Zeman, Daniel, 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. Vancouver, Canada: Association for Computational Linguistics, pp. 1–19.