Strategies for the Annotation of Pronominalised Locatives in Turkic Universal Dependency Treebanks

Jonathan Washington[1], Çağrı Çöltekin[2], Furkan Akkurt[3], Bermet Chontaeva[2], Soudabeh Eslami[2], Gulnura Jumalieva[4], Aida Kasieva[4], Aslı Kuzgun, Büşra Marşan[5], Chihiro Taguchi[6]

[1]Swarthmore College, [2]University of Tübingen, [3]Boğaziçi University, [4]Kyrgyz-Turkish Manas University, [5]Stanford University, [6]University of Notre Dame,
jwashin1@swarthmore.edu,cagri.coeltekin@uni-tuebingen.de,furkan.akkurt@bogazici.edu.tr,
{bermet.chontaeva,soudabeh.eslami}@student.uni-tuebingen.de,{gulnur.jumalieva,
aida.kasieva}@manas.edu.kg,kuzgunasli@gmail.com,
busra@stanford.edu,ctaguchi@nd.edu

Abstract

As part of our efforts to develop unified Universal Dependencies (UD) guidelines for Turkic languages, we evaluate multiple approaches to a difficult morphosyntactic phenomenon, pronominal locative expressions formed by a suffix *-ki*. These forms result in multiple syntactic words, with potentially conflicting morphological features, and participating in different dependency relations. We describe multiple approaches to the problem in current (and upcoming) Turkic UD treebanks, and show that none of them offers a solution that satisfies a number of constraints we consider (including constraints imposed by UD guidelines). This calls for a compromise with the 'least damage' that should be adopted by most, if not all, Turkic treebanks. Our discussion of the phenomenon and various annotation approaches may also help treebanking efforts for other languages or language families with similar constructions.

Keywords: Turkic languages, Universal Dependencies, treebanks

1. Introduction

As the number of treebanks for a single language or a language family in the Universal Dependencies (UD) repository[1] grows, consistent annotations become a concern (Gamba and Zeman, 2023a,b; Zeldes and Schneider, 2023). We report on one issue that is part of ongoing efforts to unify Universal Dependencies (UD) treebanks for Turkic languages, currently numbering at 16 in 8 different UD languages. Issues regarding the consistency of UD annotation of Turkic languages have been reported in earlier studies (Tyers et al., 2017; Türk et al., 2019; Çöltekin et al., 2022), with the main consensus being the need for more unified and consistent annotations across treebanks.

In this paper, we examine one selected issue in depth—namely, that of *-ki*, which attaches to nouns in the genitive and locative case. With locative nouns, it forms either attributive expressions or pronominals, while with genitive nouns, the result is always a pronominal expression.[2] As explained in detail in §2, how to appropriately annotate these pronominal forms is unclear and problematic with the present UD guidelines. As a result, the current Turkic treebanks adopt different approaches to annotating this construction. Divergence also exists within different treebanks of the same language.

We believe that the discussion of this linguistic phenomenon is likely to increase the consistency of current treebanks, help researchers creating new treebanks for Turkic languages (and others facing similar issues), and may result in improvements to the general UD guidelines by highlighting issues that are not well addressed in the current guidelines.

In this paper, we provide background information on the issue of pronominalised locatives (§2), discuss in depth several possibilities for the annotation of pronominalised locatives in Turkic languages (§3), summarise these approaches (§4), and conclude (§5). While a recommendation for a preferred approach is not put forth, a potential compromise is identified.

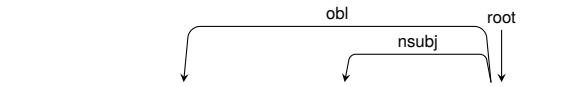
2. The issue of pronominalised locatives

In Turkic languages, locative forms of nominals (e.g., nouns, pronouns, and proper nouns) function as a locative adjunct/modifier to the head of an embedded or root clause, as in Figure 1.

Locatives cannot modify nouns on their own. One common strategy to use locatives attributively as a modifier to a noun is with the addition of the

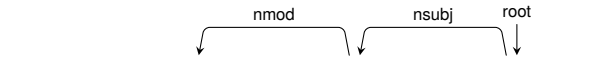
[1]See Appendix A for information on current and upcoming Turkic UD treebanks.

[2]Here, we only focus on the more varied, locative version. The outcome of the present discussion is likely to inform the issue of the annotation of genitives as well.



Turkish:	*oda-da*	*çocuk-lar*	*uyu-du-lar*
Azerbaijani:	*otaq-da*	*uşaq-lar*	*yuxla-dı-lar*
Kyrgyz:	*бөлмө-дө*	*бал-дар*	*укта-ды*
Tatar:	*бүлмә-дә*	*бала-лар*	*йокла-ды*
	room-LOC	child-PL	sleep-PST(-PL)
Lemma:	oda/otaq	çocuk/uşaq	uyu/yuxla
	/бөлмө/бүлмә	/бала/бала	/укта/йокла
POS:	NOUN	NOUN	VERB
Case:	Loc	Nom	-

Figure 1: A sentence containing an attributive locative; English translation: “Children slept in the room.”



Turkish:	*oda-da-ki*	*çocuk-lar*	*uyu-du-lar*
Azerbaijani:	*otaq-da-kı*	*uşaq-lar*	*yuxla-dı-lar*
Kyrgyz:	*бөлмө-дө-гү*	*бал-дар*	*укта-ды*
Tatar:	*бүлмә-дә-гә*	*бала-лар*	*йокла-ды*
	room-LOC-ATTR	child-PL	sleep-PST(-PL)
Lemma:	oda/otaq	çocuk/uşaq	uyu/yuxla
	/бөлмө/бүлмә	/бала/бала	/укта/йокла
POS:	NOUN	NOUN	VERB
Case:	Loc	Nom	-

Figure 2: A sentence containing an attributive locative; English translation: “The children in the room fell asleep.”

morpheme *-ki*,[3] as in Figure 2.[4]

When a locative is used attributively in this way, we opt to annotate it as `nmod` or `nmod:loc`,[5] since it is a nominal dependent (with a noun POS and lemma) of a nominal, just as in the semantically equivalent English sentence. A disadvantage of this approach is that the `Case` feature remains `Loc` and the *-ki* morpheme is not treated separately. However, the structure is recoverable, as these constructions are unique (in each language where it occurs) as the only time a locative `nmod` dependent is found.

As with other attributive expressions in Turkic languages—including adjectives per Krejci and Glass (2015) and verbal adjectives per Washington et al. (2022)—these attributive locative expressions may be used nominally, as a sort of pronominal.[6] We consider this a form of syntactic derivation.[7] For example, the sentence in Figure 2 may be expressed without the noun head of the *-ki* bearing form, with any morphology normally found there being found on the dependent *-ki* bearing form instead, as in Figure 3.

Turkish:	*oda-da-ki-ler*	*uyu-du-lar*
Azerbaijani:	*otaq-da-kı-lar*	*yuxla-dı-lar*
Kyrgyz:	*бөлмө-дө-гү-лөр*	*укта-ды*
Tatar:	*бүлмә-дә-гә-ләр*	*йокла-ды*
	room-LOC-ATTR-PL	sleep-PST(-PL)
Lemma:	oda/otaq	uyu/yuxla
	/бөлмө/бүлмә	/укта/йокла
POS:	NOUN	VERB

Figure 3: A sentence containing a pronominalised locative; English translation “The ones in the room slept.”

The resulting pronominal is formed from one noun (in this case, the room), and refers to another referent (such as the children, in this case). Several problems arise from this type of construction since there are two semantic referents (in this case, the room and the ones sleeping there) represented by a single token. Each referent has its own case, number, possessor, and other nominal features expressed through the morphology. While the locative referent still has an `nmod` relation to the other referent and contributes the `Lemma` on which the form is built, it is the other referent that has external relations: in this example, the pronominal is `nsubj` of the `root`. Conversely, the noun would be the head of any adjectival or other dependents. For example, if we add *büyük* ‘big’ to the Turkish sentence, *büyük odadakiler* has two hypothetical dependency interpretations: (1) ‘the ones in the big room’ (*büyük* ‘big’ modifying *oda* ‘room’), which is the correct interpretation, and (2) ‘the big ones in the room’ (*büyük* modifying *odadakiler* ‘the ones in the room’) is not a possible interpretation. Any solution to annotation that considers the word as a single syntactic unit cannot

[3] In many Turkic languages this has phonologically reduced, e.g. to *-kI* (Azerbaijani) or *-GI* (Kyrgyz, Tatar).

[4] Turkish, Azerbaijani, Kyrgyz, and Tatar are presented as they are the Turkic languages whose UD annotation is currently being considered by the authors.

[5] These two approaches are both acceptable in our opinion, although the latter is more specific and may make identification of this construction easier, for example in an information extraction task.

[6] By ‘pronominal’, we mean that the resulting form is not a nominal but stands in for one. For example, in Turkish ***büyükleri*** *beğendim* ‘I liked **the big ones**’, the derived form of the adjective *büyük* ‘big’ has nominal morphology and refers to an unmentioned nominal. See Göksel and Kerslake (2005, p.246) for a detailed discussion.

[7] I.e., this is a productive process that occurs in the syntax. This is not to be confused with lexical derivation, which is a historical and often not fully productive process and is usually opaque to syntax. Multiple opinions exist as to the specific mechanism by which this pronominalisation operates: through ellipsis of a nominal head, through a null-headed DP, through syntactic transformations, or otherwise.

distinguish these syntactic dependencies. Moreover, such an annotation strategy implies the latter structure, where *büyük* modifies the entire token *odadakiler*.

In an ideal solution to annotation, all morphological and syntactic information about the two participants would be recoverable.

To further complicate matters, the *-ki* morpheme can be attached to the same word multiple times. Although forms with multiple *-ki* morphemes can be difficult to interpret and rare in real-world usage, there is no principled limit for the number of *-ki* morphemes that can be attached to a noun. For example, to refer to 'glasses in the cupboard in the room', we could use the Turkish expression *oda-da-**ki**-nde-**ki**-ler* 'the ones in the one in the room'. Except cognitive load, there is nothing stopping a speaker to add another *-de-ki* to refer to the drinks inside the glasses. Although we will limit our discussion to forms with a single *-ki* morpheme, the ideal solution should also work well for words with multiple occurrences of the morpheme.

In summary, considering the pronominal forms created with the morpheme *-ki* as single syntactic words results in two major issues (see Çöltekin, 2016, for an earlier discussion):

- It violates the *lexical integrity principle* (Haspelmath and Sims, 2010, p.203) since the syntactic dependencies refer to parts of words.
- It also results in conflicting morphological features. For example, in the example in Figure 3, 'room' is singular, while the resulting pronominal refers to multiple people in the room.

The following sections discuss various ways we see as possible approaches to annotating these nominalised constructions in UD.

3. Possible Approaches

Here we demonstrate four possible approaches to the annotation of pronominalised locative forms and discuss advantages and disadvantages of each: keeping a single token (3.1), using layered features (3.2), splitting the token before *-ki* (3.3), and splitting the token after *-ki* (3.4).

We will use the Turkish sentence *Bardak dolabında**ki**lerim düştüler* '**The one**s of mine on the cup cabinet fell' to illustrate how different approaches handle these forms.

The pronominal in this sentence refers to a group of items, e.g., glasses, papers, etc. This example was chosen because there are different number, case, and possession features morphologically indicated for each of the two referents of the pronominalised locative token (the referent of the noun it is formed around and the referent of the pronominal it comprises). An alternative version of this sentence with an independent noun modified by a *-ki* bearing form is provided with annotation in Figure 4 for reference.

3.1. No segmentation

The first option is to have **no segmentation** of the word *dolabındakilerim* 'the ones of mine on its cabinet', as presented in Figure 5.

The advantage of this choice is practical: subword segmentation is a non-trivial task, and avoiding it will help make automated segmentation more precise, especially in low-resource settings. On the other hand, it is not clear what values to assign to the `Number`, `Person[psor]`, or `Person` categories, since the values for both referents of the token *dolabındakilerim* are present: the noun is singular, locative, and has a third-person possessor, while the resulting pronominal is plural, nominative, and has a first-person (plural) possessor.[8] This choice additionally fails to capture several aspects of the dependencies in this sentence:

- that there are two referents of the form: a noun and a pronominal;
- that there is a relationship between the form's two referents;
- that the first noun token in the sentence is a possessor `nmod` of the form's first referent (the noun) and not the second (the pronominal); and
- that the second referent of the form (the pronominal) and not the first (the noun) is the `nsubj` of the `root`.

Current treebanks employing a no-segmentation approach in Turkish[9] assume an analysis of elision and use the concept of promotion (whereby a normally dependent function word is 'promoted' to the syntactic function that an elided head would normally have)[10] to annotate dependencies. In our example *oda-da-ki-ler* 'the ones in the room', this approach considers the head word *çocuk-lar* 'children' to be elided.[11] Hence, its dependent *odadaki* is promoted to

[8] All other combinations are also possible in other contexts; for example, *dolaplarındakilerim*, *dolaplarındakim*, or *dolabındakim*.

[9] I.e., Penn (Cesur et al., 2023a), KeNet (Kuzgun et al., 2023b), FrameNet (Cesur et al., 2023b), Tourism (Kuzgun et al., 2023a), Atis (Köse and Yıldız, 2023).

[10] Per `https://universaldependencies.org/u/overview/syntax.html`.

[11] Unlike the English translation where the pronoun *one* still occupies the head of the construction.

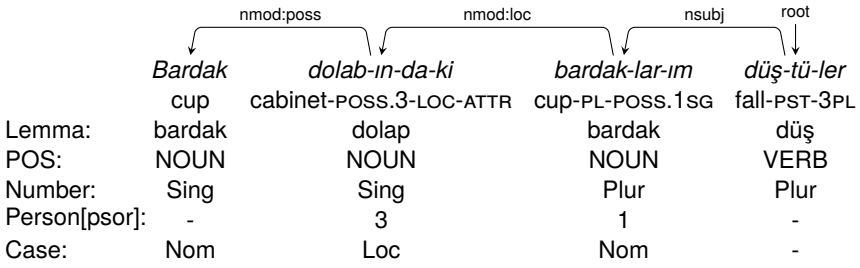


Figure 4: Analysis of a sentence comparable to the reference sentence but with a full noun phrase.

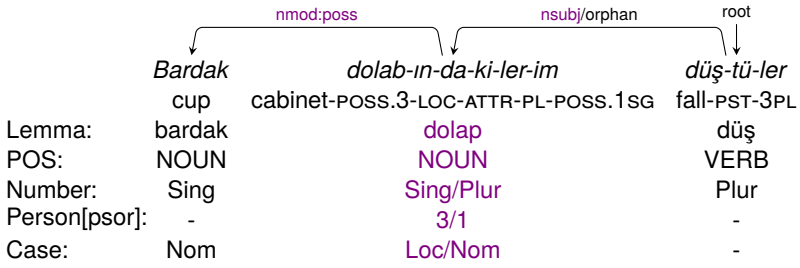


Figure 5: Analyzing *-ki* with no segmentation.

`nsubj`. According to this approach as taken in these treebanks, the *-ki* bearing form in the example in Figure 5 is an `nsubj` dependent of the verb.

Using a `Case=Loc` feature (as opposed to `Case=Nom`) with, for example, an `nsubj` dependent could clarify that this pronominal has some special status. However, a naïve downstream interpretation may understand this to be, in this example, an oblique (locative-marked) subject as opposed to a pronominal locative, especially given that the lemma is that of the attributive word (here, *dolap* 'cabinet') as opposed to the referent to which the morphology and head dependency refer (here, the pronominal referring to e.g., *bardak* 'cup'). Therefore, one option is to use the `orphan` tag when the *-ki* word is pronominal, shown as an option in Figure 5. The `orphan` relation is traditionally used in cases of head ellipsis where there is a remnant nominal that must attach to a head that it would not normally attach to. This approach solves the issue with misleading annotations; however, the `orphan` analysis is not informative. Furthermore, the issues with multiple `Number`, `Person[psor]`, and `Case` features that need to be assigned to the form *odadakiler* remain.

Another option is to introduce a new case feature for attributive and pronominal locative, such as `AttrLoc`. In pronominal uses, as shown in Figure 6, it would then be clear that this structure is not, for example, an oblique subject form of the lemma, but a pronominalised form of an attributive locative formed around the lemma. This at first appears to solve the problem having multiple case features, but it is still not clear how to annotate the second case feature (which can be any of the cases available in a given Turkic language). The problems of multiple number features and possessor person features also remain.

3.2. Layered features

An approach that would allow for annotation of different morphological features for the two referents of a pronominalised locative token is to use layered features.

While not currently used in this way in UD, layered features enable us to annotate more than one value on a feature key. Some Turkic treebanks have already employed layered features to annotate possessive marker on a nominal (cf. *dolab-ın-da-ki* and *bardak-lar-ım* in Figure 4, where `psor` in the brackets specifies that the `Person` key refers to the `Person` feature of the possessor). By extending their usage, it is possible to use layered features to specify which stem a feature key is referring to. The application of this approach on the example sentence is shown in Figure 7.

Advantages of this approach are that (i) we can annotate multiple features sharing the same key without splitting the word, (ii) layers can be recursively applied, (iii) layered features can be applied to languages without a derivational morpheme like *-ki* (e.g., some Tungusic, Quechuan, and Dargin languages), and (iv) it is compatible with the hierarchical annotation of morphology in UniMorph 4.0 (Batsuren et al., 2022).

This approach, however, fails to solve the dependency relation issues presented by having a single token: it is not clear which subword token is

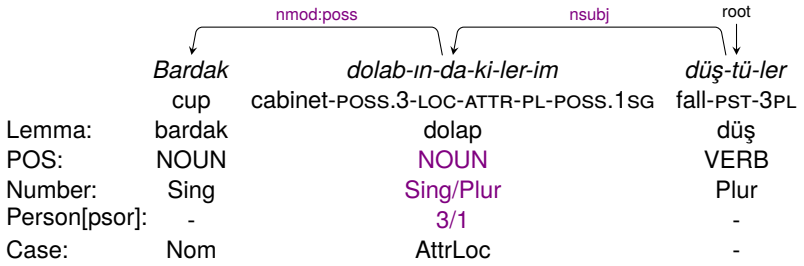


Figure 6: Analyzing *-ki* with no segmentation, with an `AttrLoc` case feature.

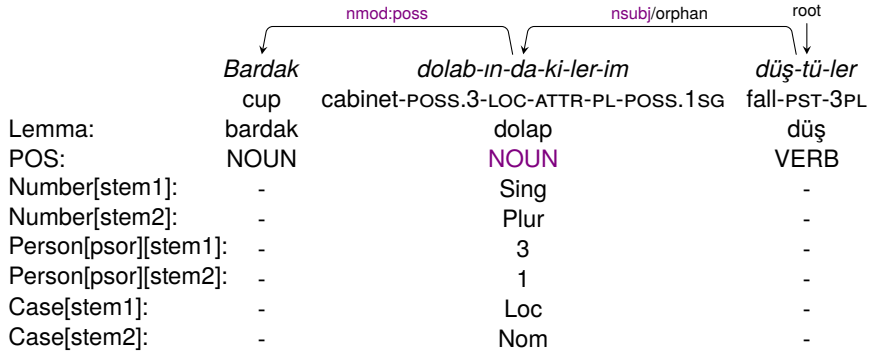


Figure 7: Analyzing *-ki* with no segmentation, using (extended) layered features.

the 'head', and which is the actual referent of the external dependency relation. There is also still only one POS.

In summary, there is a strong indication that the pronominal formed by *-ki* contains multiple syntactic words.

3.3. Splitting before *-ki*

Segmentation of the pronominalised forms solves the problems with conflicting features and dependencies, as well as the non-informativeness of the `orphan` relation. We consider two different ways (or locations) for segmenting these forms. The first option (Figure 8), which is used in some of the current treebanks (e.g., Türk et al., 2019; Marşan et al., 2022), considers the *-ki* morpheme as part of the second token.

This approach allows retaining all linguistic information packed in the *-ki* bearing forms:

- There are two referents: The possessor of the cup cabinet (third person singular) and the possessor of the items in the cabinet (first person singular). Both are clearly annotated in morphological features and POS tags in two subword tokens.
- The relationship between the two subwords is established (`nmod`, second subword being the head), and the external relationships between the *-ki* bearing form and other element(s) in the sentence are clear (the second subword being an `nsubj` dependent).
- The first subword can be annotated as taking part in other syntactic phenomena, such as compounding, independently of the full token. In our example here, the compound *bardak dolabı* is independent of (although a part of) the pronominal that is formed with *-ki*. Splitting the *-ki* bearing form into subwords allows illustrating such constructions more clearly.

In addition to enabling annotation of all morphological features and dependency relations, splitting before *-ki* prevents ending up with null morphemes (discussed in detail in §3.4). There are two disadvantages to this approach. Firstly, the current UD guidelines are not very supportive of subword tokenization, so this approach diverges from the UD framework to some extent. Secondly, due to the additional complexity, this approach can introduce noise or learnability issues for less sophisticated systems like shallow parsers.

3.4. Splitting after *-ki*

An alternative segmentation approach segments pronominalised locatives after *-ki*, as shown in Figure 9.

When splitting before *-ki*, the *-ki* morpheme is considered part of the pronominal 'word' (i.e., the part of the token representing the second referent). This can be viewed as inconsistent with the attributive use of *-ki*, where—regardless of whether or not *-ki* is best treated as an independent token—it is clear that *-ki* is not the lemma to which the

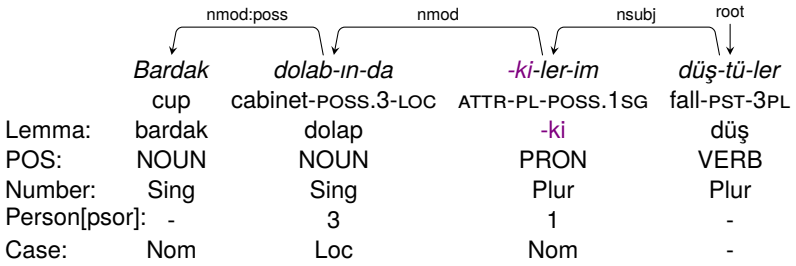


Figure 8: Possible analysis segmenting before *-ki*.

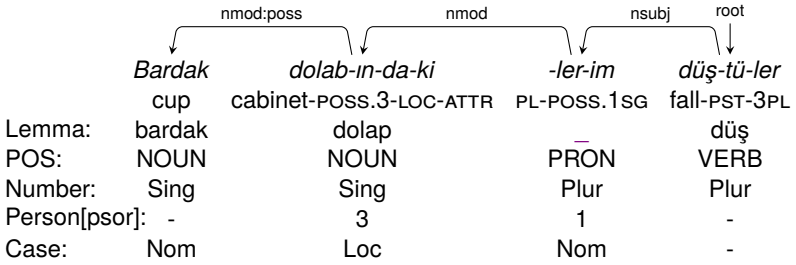


Figure 9: Possible analysis segmenting after *-ki*.

second set of morphological features belong. For example, in the annotation of the attributive use of *-ki* in Figure 4, the noun head of the *-ki* bearing form has the lemma *bardak*. However, in the annotation of an equivalent sentence with that noun absent and its morphology instead associated with the *-ki* bearing form, such as that in Figure 8, the pronoun head of the second referent (which could still be understood to refer to *bardak*), is now *-ki* according to the split-before approach. In other words, the *-ki* is associated with a different token in these two examples—and more broadly, in these two constructions: in an attributive construction, *-ki* is associated with the first participant, and in an equivalent pronominal construction, *-ki* is associated with the second participant.

The approach of splitting after *-ki*, then, is a way to avoid what might be seen as an inconsistency that arises when splitting before *-ki*. By segmenting pronominalised locatives immediately after *-ki*, the *-ki* morpheme remains with the first of the two tokens (the dependent and not the head) whether attributive or pronominal. This also unifies these two uses of *-ki* as a single phenomenon, with the addition of the phenomenon that allows the head noun to be absent in pronominal *-ki* forms.

A major problem with this approach is that it requires an empty lemma, as well as an empty form when there are no additional affixes after *-ki*. Empty lemmas and forms are not allowed according to UD v2 annotation guidelines. While it would be possible not to annotate a second token (the pronoun / second referent) if it were empty, that would reduce the consistency of this approach, and still leaves the issue of having an empty lemma. Furthermore, as with segmenting before *-ki*, there may be limitations for less sophisticated automated annotation systems, although it is possible that systems capable of segmenting words into subword units would be able to handle one approach more easily than the other—an area for future investigation. Lastly, treating attributive and pronominal locatives uniformly may go against a generative syntax analysis of these two uses, where the attributive locative form is an ordinary member of the phrase (DP) containing the head noun, whereas the pronominal locative is cast directly into a DP with the accompanying morphology and has fewer layers between the two phrases.

3.5. Splitting after *-ki* with fallback

One problem with splitting after *-ki* is that null nodes would result in situations where there is no inflection, as in the sentence *Bardak dolabındaki düştü* 'The one on the cup cabinet fell'. This problem could be avoided with a fallback in such cases.

One option is to fall back to an `orphan` analysis, per Figure 10, signalling to downstream tasks that information is missing (specifically an elided [pronominal] element). Using the `orphan` relation has the disadvantages discussed in §3.1: it is not informative, and does not allow for annotation of multiple relations (although implies them) or multiple sets of features. However, examples of pronominalised locatives are not very frequent in existing corpora, and examples of pronominalised locatives with no further inflection are quite rare, so this approach would not result in excessive use of the `orphan` relation.

To include the elided information, enhanced dependencies may be used, as in Figure 11. En-

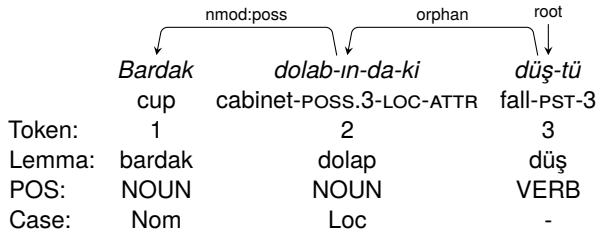


Figure 10: Possible analysis segmenting after *-ki* with no morphology, with `orphan` fallback.

hanced dependencies are explicitly designed to present null nodes in cases of elision.[12] Use of enhanced dependencies has some drawbacks. If annotated even for just one example, the entire corpus needs to have enhanced dependencies annotated. Furthermore, most parsers, querying tools, and other applications of UD lack support for enhanced dependencies and ignore them. However, this approach does preserve the information lost in the accompanying standard dependency analysis.

4. Summary of approaches

The approaches described in Section 3, and their advantages and disadvantages are summarised in Table 1.

The first approach discussed, no-segmentation (§3.1), has the benefit of ease of tokenization. Even though state-of-the-art parsers may be successful in segmenting words into subword units, not having to split words has a clear advantage, especially in low-resource scenarios.[14] It also avoids empty word forms and empty lemmas that some of the approaches postulate. However, it fails to represent multiple sets of morphological features, and it does not allow a correct interpretation of the dependency relations the word participates in. Specifically, annotating in this way results in a situation where it is unclear which of the token's referents is the modifier of another head. Possible ways to remove the ambiguity would be to use the `orphan` relation (second row of Table 1) or an `AttrLoc` value for the case feature (third row of Table 1), both of which allow for differentiation of pronominalised locatives from other dependents with a similar relation to the head. However, `orphan` does not include any information regarding the syntactic function of the word in the sentence. With or without the `orphan` relation or an `AttrLoc` case feature, the no-segmentation approach does not resolve the issue of multiple, potentially conflicting sets of morphological features assigned to a single syntactic word.

A possible solution (described in §3.2) that allows expressing multiple sets of morphological features is to make use of layered features as exemplified in Figure 7. Although this uses the UD layered features in an unorthodox way,[15] it enables specification of multiple sets of morphological features, and, with the use of the `orphan` relation, pronominalised locatives can also be differentiated from other dependents with a similar relation to their head. However, as noted earlier, it does not allow identifying the dependency relations correctly. It still leaves it unclear which part of the word is modified by a modifier, and which part is a modifier to another head. Another downside is, perhaps, the complexity: such feature sets and relations are likely to be difficult to learn for parsers, and the treebank queries for relevant features/structures are likely to be misled or miss the relevant items due to the idiosyncratic nature of the annotations.

Both segmentation options resolve the main concerns with the pronominal construction: the appropriate features are easily assigned to each syntactic word, and the dependents can modify the correct syntactic word without ambiguity. The relation between the pronominal and its head is also clearer. The disadvantage of splitting before *-ki* (§3.3) is the inconsistency with the attributive use. This approach suggests either splitting *-ki* in attributive usage without any clear motivation—in which case it is still not the lemma of the modified noun's morphological features as in the pronominal treatment—or treating attributive and pronominal cases differently.[16] The disadvantage of splitting after *-ki* (§3.4) is the introduction of empty lemmas, and empty forms when no further affixes are attached after *-ki*. Since empty forms are not allowed in the current basic UD dependencies, this approach would require a substantial modification to the UD guidelines. Splitting after *-ki* with fallback (§3.5) solves the issue of empty lemmas but requires the use of enhanced dependencies

[12] Per https://universaldependencies.org/v2/enhanced.html.

[13] Empty forms and lemmas would only occur in enhanced dependencies annotation, where they are permissible.

[14] We intend to investigate this empirical question in future research.

[15] E.g., introducing multi-dimensional layers, and layers indexed by ordinals.

[16] Which may also result in difficulties with the automated processing.

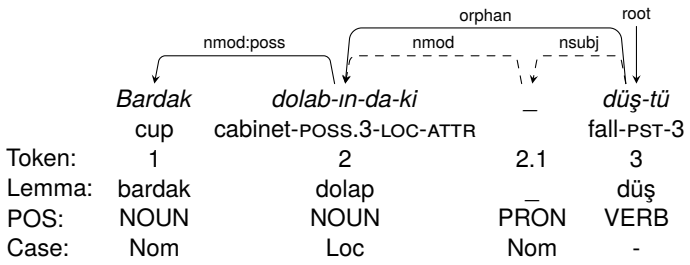


Figure 11: Possible analysis segmenting after *-ki* with no morphology, with enhanced dependencies fallback.

Approach	No empty forms	No empty lemmas	2 sets of features	Deprels for 2 referents	Consistent with attributive use	Easy querying	No need for subword-aware parser
No-segmentation	✔	✔	✘	✘	✘	✔	✔
`orphan` relation	✔	✔	✘	✘	✘	✔	✔
`AttrLoc` feature	✔	✔	✘	✘	✘	✔	✔
Layered features	✔	✔	✔	✘	✘	✘	✔
Splitting before *-ki*	✔	✔	✔	✔	✘	✔	✘
Splitting after *-ki*	✘/✔	✘	✔	✔	✔	✔	✘
Splitting after, enhanced dependencies fallback	(✔)[13]	(✔)[13]	✔	✔	✔	✘	✘

Table 1: A summary of the advantages and disadvantages in the discussed approaches.

framework, which introduces a new set of challenges including compatibility issues for existing UD tools.

5. Concluding remarks

The authors currently consider splitting pronominalised locatives before *-ki* a best compromise, and recommend this for annotation of Turkic treebanks, although with a caveat.

While the authors agree with one another that segmentation is needed to properly capture these constructions, opinions differ as to which approach is ideal. Proponents of splitting the pronominalised locative before *-ki* do not believe that it is a problem for the approach to be inconsistent with the treatment of the attributive locative due to a generative syntax view that they are in fact distinct. Proponents of splitting the pronominalised locative after *-ki* realise that it would take a major change to current UD guidelines for this approach to be viable, and while finding splitting before *-ki* somewhat unsatisfactory, accept that it may be the current best compromise.

The issue of pronominalised locatives is just one of many specific issues where consistent UD annotation guidelines are needed for Turkic languages. This issue is also relevant to the UD (and UniDive) community at large. By bringing awareness to this issue and discussing it in depth, we hope that new annotation projects for languages with similar phenomena will be eased, and that our efforts will lead to improved overall quality of corpora and annotation guidelines.

6. Acknowledgements

We are grateful to Gülnara Karasawa for her help with creating the Tatar dataset for this study.

We thank UniDive, the COST Action CA21167, and Istanbul Technical University for supporting the organization of the workshop in September 2023 that made the present work and future collaboration towards a unified annotation of Turkic UD treebanks possible. We also thank all participants of the workshop for fruitful discussion and suggestions.

7. Bibliographical References

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam

Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Özlem Çetinoğlu and Çağrı Çöltekin. 2019. Challenges of annotating a code-switching treebank. In *Proceedings of the 18th international workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 82–90.

Özlem Çetinoğlu and Çağrı Çöltekin. 2022. Two languages, one treebank: building a Turkish–German code-switching treebank and its challenges. *Language Resources and Evaluation*, pages 1–35.

Çağrı Çöltekin. 2015. A grammar-book treebank of Turkish. In *Proceedings of the 14th workshop on Treebanks and Linguistic Theories (TLT 14)*, pages 35–49.

Çağrı Çöltekin, A Doğruöz, and Özlem Çetinoğlu. 2022. Resources for Turkish natural language processing: A critical survey. *Language Resources and Evaluation*.

Mehmet Oguz Derin and Takahiro Harada. 2021. Universal Dependencies for Old Turkish. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 129–141, Sofia, Bulgaria. Association for Computational Linguistics.

Marhaba Eli, Weinila Mushajiang, Tuergen Yibulayin, Kahaerjiang Abiderexiti, and Yan Liu. 2016. Universal dependencies for Uyghur. In *Proceedings of the Third International Workshop on Worldwide Language Service Infrastructure and Second Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies (WLSI/OIAF4HLT2016)*, pages 44–50, Osaka, Japan. The COLING 2016 Organizing Committee.

Federica Gamba and Daniel Zeman. 2023a. Latin morphology through the centuries: Ensuring consistency for better language processing. In *Proceedings of the Ancient Language Processing Workshop*, pages 59–67, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Federica Gamba and Daniel Zeman. 2023b. Universalising Latin Universal Dependencies: a harmonisation of Latin treebanks in UD. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 7–16, Washington, D.C. Association for Computational Linguistics.

Aslı Göksel and Celia Kerslake. 2005. *Turkish: A Comprehensive Grammar*. London: Routledge.

Martin Haspelmath and Andrea D. Sims. 2010. *Understanding Morphology*, second edition. Understanding Language. Taylor & Francis.

Aida Kasieva, Gulnura Dzhumalieva, Anna Thompson, Murat Jumashev, Bermet Chontaeva, and Jonathan Washington. 2023. Issues of Kyrgyz syntactic annotation within the Universal Dependencies framework. In *Proceedings of the XI International Conference on Computer Processing of Turkic Languages (TurkLang 2023)*.

Bonnie Krejci and Lelia Glass. 2015. The Kazakh noun/adjective distinction. In *Proc. of the 9th Workshop on Altaic Formal Linguistics (WAFL9)*, pages 47–58, Cambridge, MA. MITWPL.

Aibek Makazhanov, Aitolkyn Sultangazina, Olzhas Makhambetov, and Zhandos Yessenbayev. 2015a. Syntactic annotation of Kazakh: Following the universal dependencies guidelines. a report. In *Proceedings of the 3rd International Conference on Computer Processing in Turkic Languages (TurkLang 2015)*, pages 338–350.

Aibek Makazhanov, Aitolkyn Sultangazina, Olzhas Makhambetov, and Zhandos Yessenbayev. 2015b. Syntactic annotation of Kazakh:

Following the universal dependencies guidelines. a report. In *3rd International Conference on Turkic Languages Processing, (TurkLang 2015)*, pages 338–350.

Büşra Marşan, Salih Furkan Akkurt, Muhammet Şen, Merve Gürbüz, Onur Güngör, Şaziye Betül Özateş, Suzan Üsküdarlı, Arzucan Özgür, Tunga Güngör, and Balkız Öztürk. 2022. Enhancements to the boun treebank reflecting the agglutinative nature of turkish. In *The Proceedings of the ALTNLP2022 The International Conference and workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing*, pages 71–80.

Tatiana Merzhevich and Fabrício Ferraz Gerardi. 2022. Introducing YakuToolkit. Yakut treebank and morphological analyzer. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 185–188, Marseille, France. European Language Resources Association.

Umut Sulubacak, Memduh Gokirmak, Francis Tyers, Çağrı Çöltekin, Joakim Nivre, and Gülşen Eryiğit. 2016. Universal Dependencies for Turkish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3444–3454, Osaka, Japan. The COLING 2016 Organizing Committee.

Utku Türk, Furkan Atmaca, Şaziye Betül Özateş, Balkız Öztürk Başaran, Tunga Güngör, and Arzucan Özgür. 2019. Improving the annotations in the Turkish Universal Dependency treebank. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 108–115, Paris, France. Association for Computational Linguistics.

Utku Türk, Furkan Atmaca, Şaziye Betül Özateş, Gözde Berk, Seyyit Talha Bedir, Abdullatif Köksal, Balkız Öztürk Başaran, Tunga Güngör, and Arzucan Özgür. 2022. Resources for turkish dependency parsing: Introducing the boun treebank and the boat annotation tool. *Language Resources and Evaluation*, pages 1–49.

Francis Tyers and Jonathan Washington. 2015. Towards a free/open-source universal-dependency treebank for Kazakh. In *Proceedings of the 3rd International Conference on Computer Processing in Turkic Languages (TurkLang 2015)*, pages 276–289.

Francis Tyers, Jonathan Washington, Çağrı Çöltekin, and Aibek Makazhanov. 2017. An assessment of Universal Dependency annotation guidelines for Turkic languages. In *5th International Conference on Turkic Language Processing (TURKLANG 2017)*, pages 356–377.

Jonathan N. Washington, Francis M. Tyers, and Ilnar Salimzianov. 2022. Non-finite verb forms in Turkic exhibit syncretism, not multifunctionality. *Folia Linguistica*, 56(3):693–742.

Jonathan North Washington and Francis Morton Tyers. 2019. Delineating Turkic non-finite verb forms by syntactic function. In *Proceedings of the Workshop on Turkic and Languages in Contact with Turkic*, volume 4, pages 115–129.

Amir Zeldes and Nathan Schneider. 2023. Are UD treebanks getting more consistent? a report card for English UD. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 58–64, Washington, D.C. Association for Computational Linguistics.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, 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 de Paiva, 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 Mendonça, 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. Association for Computational Linguistics.

Çağrı Çöltekin. 2016. (When) do we need inflectional groups? In *Proceedings of The First International Conference on Turkic Computational Linguistics*.

8. Language Resource References

Ibrahim Benli. 2023. *Kyrgyz KTMU Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Cesur, Neslihan and Kuzgun, Aslı and Yıldız, Olcay Taner and Marşan, Büşra and Kara, Neslihan and Arıcan, Bilge Nas and Özçelik, Merve and Aslan, Deniz Baran. 2023a. *Turkish Penn Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Cesur, Neslihan and Kuzgun, Aslı and Yıldız, Olcay Taner and Marşan, Büşra and Kuyrukçu, Oğuzhan and Arıcan, Bilge Nas and Sanıyar, Ezgi and Kara, Neslihan and Özçelik, Merve. 2023b. *Turkish FrameNet Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Eli, Marhaba and Zeman, Daniel and Tyers, Francis. 2023. *Uyghur UDT Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Kuzgun, Aslı and Cesur, Neslihan and Yıldız, Olcay Taner and Kuyrukçu, Oğuzhan and Marşan, Büşra and Arıcan, Bilge Nas and Kara, Neslihan and Aslan, Deniz Baran and Sanıyar, Ezgi and Asmazoğlu, Cengiz. 2023a. *Turkish Tourism Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Kuzgun, Aslı and Cesur, Neslihan and Yıldız, Olcay Taner and Kuyrukçu, Oğuzhan and Yenice, Arife Betül and Arıcan, Bilge Nas and Sanıyar, Ezgi. 2023b. *Turkish Kenet Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Köse, Mehmet and Yıldız, Olcay Taner. 2023. *Turkish Atis Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Makazhanov, Aibek and Washington, Jonathan North and Tyers, Francis. 2023. *Kazakh KTB Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Marşan, Büşra and Akkurt, Salih Furkan and Türk, Utku and Atmaca, Furkan and Özateş, Şaziye Betül and Berk, Gözde and Bedir, Seyyit Talha and Köksal, Abdullatif and Başaran, Balkız Öztürk and Güngör, Tunga and Özgür, Arzucan. 2023. *Turkish BOUN Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Merzhevich, Tatiana and Gerardi, Fabrício Ferraz. 2023. *Yakut YKTDT Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Taguchi, Chihiro. 2023. *Tatar NMCTT Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Türk, Utku and Özateş, Şaziye Betül and Marşan, Büşra and Akkurt, Salih Furkan and Çöltekin, Çağrı and Cebiroğlu Eryiğit, Gülşen and Gökırmak, Memduh and Kaşıkara, Hüner and Sulubacak, Umut and Tyers, Francis. 2023. *Turkish IMST Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Uszkoreit, Hans and Macketanz, Vivien and Burchardt, Aljoscha and Harris, Kim and Marheinecke, Katrin and Petrov, Slav and Kayadelen, Tolga and Attia, Mohammed and Elkahky, Ali and Yu, Zhuoran and Pitler, Emily and Lertpradit, Saran and Cetin, Savas and Popel, Martin and Zeman, Daniel and Tyers, Francis and Çöltekin, Çağrı and Türk, Utku and Atmaca, Furkan and Özateş, Şaziye Betül and Köksal, Abdullatif and Başaran, Balkız Öztürk and Güngör, Tunga and Özgür, Arzucan. 2023. *Turkish PUD Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Çetinoğlu, Özlem and Çöltekin, Çağrı. 2023. *Turkish German SAGT Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Çöltekin, Çağrı. 2023. *Turkish GB Treebank of Universal Dependencies 2.13*. Universal Dependencies Consortium. PID http://hdl.handle.net/11234/1-5287. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A. UD Turkic Treebanks

There are currently UD treebanks for Kazakh, Kyrgyz, Tatar, Turkish, Uyghur, Yakut, and Old Turkish, and a treebank annotating sentences with Turkish-German code switching. All languages except Turkish are represented with a single treebank, while Turkish has 9 treebanks. Table 2 lists the treebanks currently released in the UD repositories as of UD version 2.13.

	sent	tok	multi	types	ltypes	pos	rel	feat
Kazakh/KTB (Tyers and Washington, 2015; Makazhanov et al., 2015a) (Makazhanov et al., 2023)	1078	10536	41	4642	2433	17	36	9
Kyrgyz/KTMU (Benli, 2023)	781	7451	0	3474	2305	13	26	8
Old Turkish/Tonqq (Derin and Harada, 2021)	20	158	0	75	2	13	19	0
Tatar/NMCTT (Taguchi, 2023)	148	2280	0	1264	843	14	28	7
Turkish/Atis (Köse and Yıldız, 2023)	5432	45907	0	2133	995	13	36	7
Turkish/BOUN (Türk et al., 2022; Marşan et al., 2022) (Marşan et al., 2023)	9761	125212	3374	37052	12649	16	46	7
Turkish/FrameNet (Cesur et al., 2023b)	2698	19223	0	8403	3905	15	30	7
Turkish/GB (Çöltekin, 2015) (Çöltekin, 2023)	2880	17177	371	5517	2074	16	42	7
Turkish/IMST (Sulubacak et al., 2016) (Türk et al., 2023)	5635	58096	1639	18541	5960	14	40	10
Turkish/Kenet (Kuzgun et al., 2023b)	18687	178658	0	49156	15343	15	34	7
Turkish/Penn (Cesur et al., 2023a)	16396	183555	0	37765	14977	15	36	9
Turkish/PUD (Zeman et al., 2017) (Uszkoreit et al., 2023)	1000	16881	346	7646	4598	16	38	4
Turkish/Tourism (Kuzgun et al., 2023a)	19830	91152	0	4961	2170	15	33	13
Turkish-German/SAGT (Çetinoğlu and Çöltekin, 2022) (Çetinoğlu and Çöltekin, 2023)	2184	37227	290	7094	3836	17	45	12
Uyghur/UDT (Eli et al., 2016) (Eli et al., 2023)	3456	40236	0	12067	2908	16	45	15
Yakut/YKTDT (Merzhevich and Ferraz Gerardi, 2022) (Merzhevich and Gerardi, 2023)	299	1460	1	688	405	14	26	6

Table 2: Basic statistics on current UD treebanks (as of UD version 2.13). *sent*: number of sentences, *tok*: number of tokens, *multi*: number of multi-word tokens, *types*: number of word types, *ltypes*: number of lemma types, *pos*: number of POS tags used, *rel*: number of dependency relations used (including language/treebank specific relations), *feat*: number of morphological features used.

Besides existing treebanks, the UD web page also reports Uzbek, Ottoman Turkish and yet another Turkish treebank in preparation. We are also aware of new treebanks in preparation for Kyrgyz (Kasieva et al., 2023), Azerbaijani and Kumyk.