Harmonizing Annotation of Turkic Postverbial Constructions: A Comparative Study of UD Treebanks

Arofat Akhundjanova
Saarland University / Saarbrücken, Germany
arak00001@stud.uni-saarland.de

Abstract

As the number of treebanks within the same language family continues to grow, the importance of establishing consistent annotation practices has become increasingly evident. In this paper, we evaluate various approaches to annotating Turkic postverbial constructions across UD treebanks. Our comparative analysis reveals that none of the existing methods fully capture the unique semantic and syntactic characteristics of these complex constructions. This underscores the need to adopt a balanced approach that can achieve broad consensus and be implemented consistently across Turkic treebanks. By examining the phenomenon and the available annotation strategies, our study aims to improve the consistency of Turkic UD treebanks and enhance their utility for cross-linguistic research.

1 Introduction

As the Universal Dependencies (UD) project (Nivre et al., 2016, 2020) continues to grow, the need for consistent annotation practices across treebanks has become increasingly evident, especially for languages within the same language family. The Turkic language family, with its rich morpho-syntactic categories and agglutinative morphology, poses unique challenges for annotation. Despite the availability of several Turkic UD treebanks, inconsistencies in annotation schemes often hinder meaningful comparisons and cross-lingual studies, highlighting the necessity for a standardized approach.

Previous studies have emphasized inconsistencies in the annotation of Turkic languages, particularly in morphological features and dependency relations (Tyers et al., 2017). These include challenges in part-of-speech (POS) tagging, morphological features (Taguchi, 2022), and pronominalized locatives (Washington et al., 2024).

The development of the first UD treebank for Uzbek and the challenges faced during annotation prompted us to investigate a specific issue: the annotation of Turkic postverbial constructions. These constructions, which pair a converb with a postverb, convey nuanced meanings related to aspect or actionality. The dual role of postverbs — functioning both as grammatical markers and as independent verbal predicates — complicates their representation within the UD framework. Ensuring consistency while accurately reflecting the unique semantic and syntactic structure of postverbial constructions is difficult.

In this paper, we evaluate multiple approaches to annotating Turkic postverbial constructions across eleven UD treebanks of seven Turkic languages, as shown in Table 1. This issue is particularly critical given the variation not only across Turkic languages but also within the treebanks of a single language.

Our analyses and suggestions contribute to improving the consistency of Turkic UD treebanks and enhancing their value for cross-linguistic research.

The remainder of this paper is organized as follows. Section 2 provides background information on Turkic postverbial constructions. Section 3 presents a detailed analysis of four annotation approaches: `adverbial clause modifier`, `clausal complement`, `auxiliary` and `compound`. Section 4 offers recommendations for standardizing annotations, and Section 5 concludes our findings with implications for future work on Turkic UD treebanks.

Proceedings of the 18th Building and Using Comparable Corpora Workshop, pages 10–17
January 20, 2025.

Treebanks	sent	tok	genre	No. of postverbial constructions
Azerbaijani-TueCL (Eslami and Çağrı Çöltekin, 2024)	109	663	grammar	∼ 4
Kazakh-KTB (Tyers and Washington, 2015)	1078	10536	news, fiction, wiki	∼ 24
Kyrgyz-KTMU (Benli, 2020)	2480	23654	news, fiction	∼ 60
Kyrgyz-TueCL (Chontaeva and Çağrı Çöltekin, 2024)	145	1001	grammar	∼ 30
Tatar-NMCTT (Taguchi et al., 2022)	148	2280	news, non-fiction	∼ 6
Turkish-BOUN (Türk et al., 2021)	9761	125212	news, non-fiction	∼ 100
Turkish-GB (Çağrı Çöltekin, 2015)	2880	17177	grammar	∼ 3
Turkish-Kenet (Kuzgun et al., 2022)	18687	178658	grammar	N/A
Turkish-Penn (Cesur et al., 2022)	16396	183555	news, non-fiction	N/A
Uyghur-UDT (Eli et al., 2016)	3456	40236	fiction	∼ 80
Uzbek-UT (Akhundjanova, 2024)	500	5850	news, fiction	∼ 70

Table 1: Eleven Turkic UD treebanks representing seven languages selected for our comparative study.

2 Turkic Postverbial Constructions

Turkic languages use verbal constructions made up of a converb followed by an auxiliary verb, also called a 'postverb' (Ağcagül, 2004) or 'postverbial constructions with auxiliary verbs' (Johanson, 2021, 36-37). In these constructions, the converb provides the main lexical meaning, while the postverb, having lost much of its original meaning, primarily carries grammatical information like person, mood and tense. It also refines the description of the action, as in Kyrgyz `kel-ip tur` (lit. 'coming stand'), which means 'to come regularly.' The postverb adopts the converb's argument structure, forming a single grammatical unit.

This structure bears similarity to Indo-European preverbal units, where a non-inflecting element precedes a verb stem, forming a unified lexical unit. Preverbs typically modify or refine the verb's lexical meaning, adding spatial, directional, or aspectual nuances. For instance, in Sanskrit `pra gacchati` (lit. 'forth goes'), the meaning is 'he goes forth' (Booij and Van Kemenade, 2003).

The following kinds of verbs can occur as the auxiliary element in postverb constructions of various Turkic languages: tur-/dur- 'stand (up)', yat-/yot-/jat- 'lie (down)', oltur-/otur-/oʻtir- 'sit (down)', kel-/kil-/gel- 'come', ket-/git- 'go', bar-/bor- 'go', al-/ol- 'take', ber-/bir-/ver- 'give', ïd-/yubor- 'send', etc (Ağcagül, 2004, 7).

Postverbs typically convey two types of functions:

1. Actional modification: Postverbs modify the actional meaning of the lexical verb by specifying qualitative or quantitative properties such as suddenness (1) and thoroughness (2) (Ağcagül, 2004, 7), as in the following examples:

(1) Uzbek

ayt-ib qoʻy-di-m
say-CONV put-PST-1SG

'I blurted out' (lit. 'saying put')

(2) Uyghur

Oq-up čïq!
read-CONV emerge.IMP

'Read from beginning to end!' (lit. 'reading emerge')

2. Phase specification: Postverbs indicate different phases of an action, including its initial or final stages, as well as its continuity (Ağcagül, 2004, 7), as illustrated in the examples below:

(3) Turkish

yaz-ıp dur-du
write-CONV stand-PST.3SG

's/he kept writing' (lit. 'writing standed')

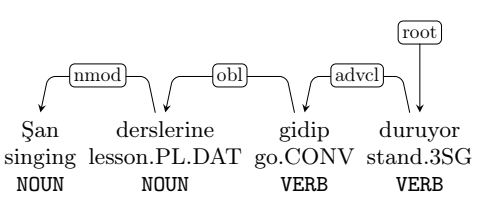
(4) Uzbek

Manzil-ga yet-ib qol-di-k
destination-DAT reach-CONV stay-PST-1PL

'We are about to reach the destination.' (lit. 'destination.to reaching (we) stayed')

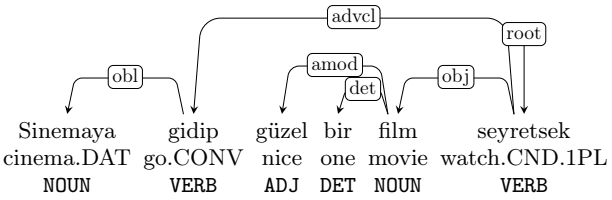
3 Existing Annotation Approaches

We examine four existing approaches to annotating Turkic postverb constructions, outlining the arguments for and against each. These approaches include treating them as `adverbial`



'S/he keeps going to singing lessons.'

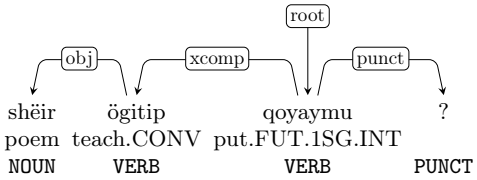
(a) Annotation for `gidip dur`.



'Let's go to the cinema and watch a nice movie.'

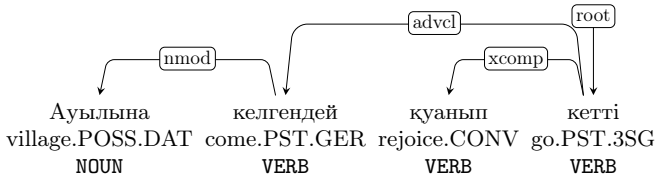
(b) Annotation for `gidip`.

Figure 1: The converb `gidip` is used in two different structures, but tagged with the same label in the Turkish-GB treebank.



'Shall I teach (you) a poem?'

(a) Annotation for `ögitip qoy`.



'S/he was happy as if s/he had come to his village.'

(b) Annotation for `қуанып кет`.

Figure 2: Annotation of the converb as `xcomp` in Uyghur and Kazakh.

`clause modifier` (3.1), `clausal complement` (3.2), `auxiliary` (3.3), and `compound` (3.4). Additionally, we find instances of mixed approaches in certain treebanks (3.5).

3.1 Adverbial clause modifier: `advcl`

One approach to addressing this issue is to annotate the converb as `advcl` and the postverb as the `head`, as shown in Figure 1a. This method has been adopted in the Turkish treebanks listed in Table 1.

However, this annotation is not ideal. The `advcl` tag is generally reserved for clauses functioning as modifiers that express temporal, causal, conditional, or similar relations. In Turkic postverb constructions, the converb does not serve as a modifier to the postverb. Instead, it forms an integral part of the verbal phrase, contributing essential lexical meaning. Annotating the converb as `advcl` misrepresents its role, inaccurately suggesting that it has a subordinate function relative to the postverb. This approach fails to capture the grammaticalized and semantically unified nature of these constructions. For comparison, see Figure 1b, which shows a true adverbial clause modifier using the same converb `gidip`, contrasted with the postverbial construction in Figure 1a.

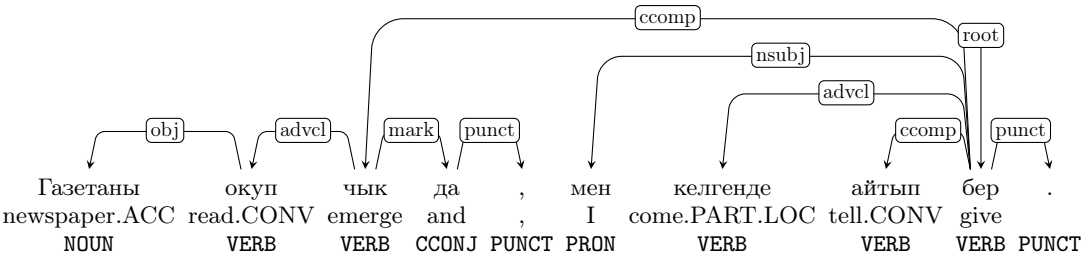
3.2 Clausal Complement: `xcomp` and `ccomp`

Another option is to tag the converb as `xcomp` (see Figure 2a for Uyghur and 2b for Kazakh) or `ccomp` (see Figure 3 for Kyrgyz) and the postverb as the head. This method is not plausible, because the two elements of postverbial constructions do not function as independent predicates, nor do they exhibit the syntactic independence typical of an `xcomp` or `ccomp` relation. In these relations, the complement clause is subordinate to the main predicate (head) and lacks its own subject, relying on an external argument for subject control. However, in postverbial constructions, the converb is not a subordinate clause but rather an integral part of a compound verb.

3.3 Auxiliary: `aux`

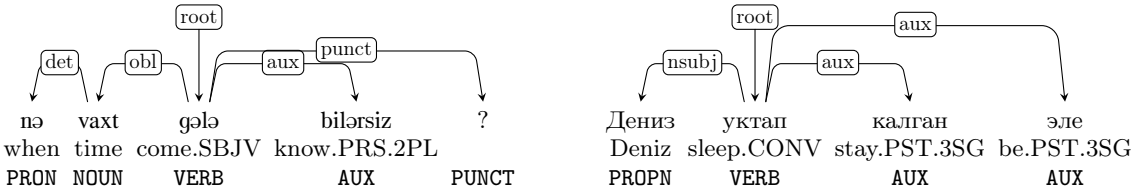
Tagging the converb as the head and the postverb as `aux` can be a reasonable approach in some contexts. See Figure 4a from Azerbaijani-TueCL, Figure 4b from Kyrgyz-TueCL and Figure 5 from Tatar-NMCTT. However, there are important considerations and potential limitations depending on the specific properties of the language.

On the one hand, the converb carries the primary lexical meaning, making it appropriate to treat it as the head. This reflects its domi-



'Read the newspaper and tell me when I come.'

Figure 3: Annotation of `окуп чык` with `advcl` and `айтып бер` with `ccomp`.



'When can you come?'

(a) Annotation for `gələ bil`.

'Deniz had fallen asleep.'

(b) Annotation for уктап кал.

Figure 4: Annotation of the converb as a head and the postverb as `aux`.

nant role in encoding the core action or state of the clause. Postverbs are often grammaticalized to indicate auxiliary-like functions, which aligns with the typical `aux` tag. Treating the postverb as `aux` captures its secondary grammatical function and reduced lexical meaning. In both Azerbaijani and Kyrgyz treebanks, this approach is applied based on the classification of auxiliaries in their respective languages. In the Azerbaijani treebank, independent verbs like `bil` 'know' and `ol` 'become' are tagged with AUX POS, and Kyrgyz-TueCL treebank has a larger list of auxiliaries: жат, кал, ал, бол, кой, кет, тур, etc. Tatar treebank (Taguchi et al., 2022) also indicates that the finite verb in grammaticalized converb constructions is marked as AUX.

On the other hand, in other Turkic languages, postverbs often retain independent, non-auxiliary uses as lexical verbs and appear as heads of their own clauses with full argument structures. For example, compare the following two Uzbek sentences:

(5) Uzbek

yomg'ir qor-ni eri-t-ib
rain snow-ACC melt-CAU-CONV
yubor-di
send-PST.3SG

'The rain melted the snow away.'

(6)

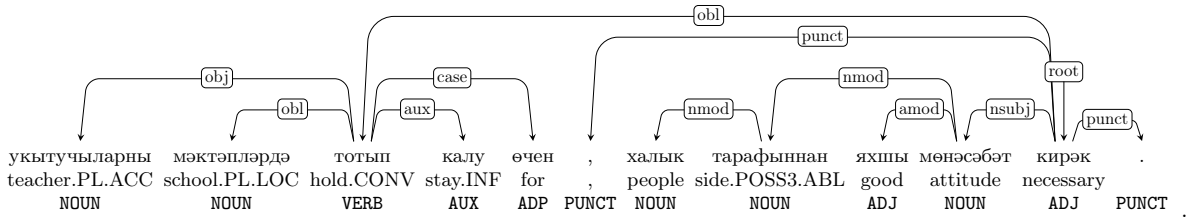
xat-ni ber-ib yubor-di
letter-ACC give-CONV send-PST.3SG

'S/he gave/sent the letter away.'

In (5), the postverb `yubordi` 'sent' marks the immediate completion of the action expressed by the converb `eritib` 'melting'. In (6), both `berib` 'giving' and `yubordi` 'sent' retain their independent meanings, and serve more like a serial verb construction (`compound:svc`). For this reason, in Uzbek, about 27 verbs that can be used as auxiliaries to form postverbial constructions are classified as `VERB`, not `AUX` and the `aux` relation is restricted to modal and copular verbs, and may not extend to aspectual or actionality markers. Hence, this approach would overload the `aux` with elements that do not fit its traditional definition.

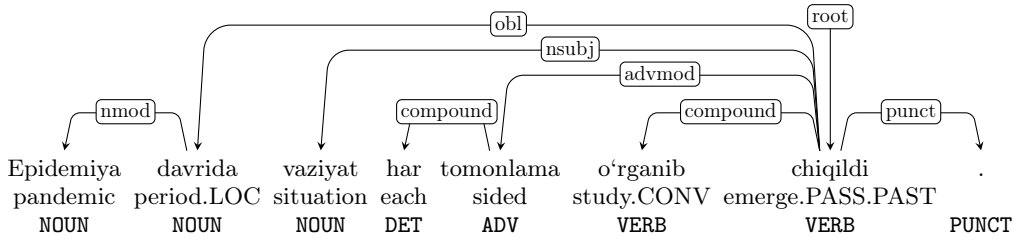
3.4 Compound

The final approach is to use a `compound` relation, as shown in the Uzbek-UT example in Figure 6. Postverb constructions are akin to compound verbs, where all elements contribute to forming a single lexical unit. However, we acknowledge that the `compound` label does not



'In order to retain teachers in schools, good public attitude is needed.'

Figure 5: Annotation of `тотып калу` as `aux`.



'During the epidemic, the situation was thoroughly studied.'

Figure 6: Annotation of `o‘rganib chiq` as `compound`.

fully reflect the postverb's desemanticized and auxiliary-like role. Tagging the converb as `compound:lvc` (light verb construction, LVC) instead could be a partially plausible option. In such verbal constructions, the verbal or nonverbal predicate provides the main semantic content like converbs in our case, while the light verb contributes grammatical information, resembling postverbs. The `compound:lvc` relation highlights the grammaticalized nature and auxiliary function of the postverb while still acknowledging the converb as the core semantic contributor. It aligns with the principle that LVCs combine a semantically strong element with a semantically weak verb.

The limitation of this approach is that Turkic postverb constructions are highly grammaticalized, often to the point where the postverb functions more like an auxiliary than a light verb. As a result, using the `compound:lvc` might not fully capture this advanced stage of grammaticalization.

3.5 Mixed Approaches

The inconsistency in annotation methods within the same language or treebank may stem from several factors.

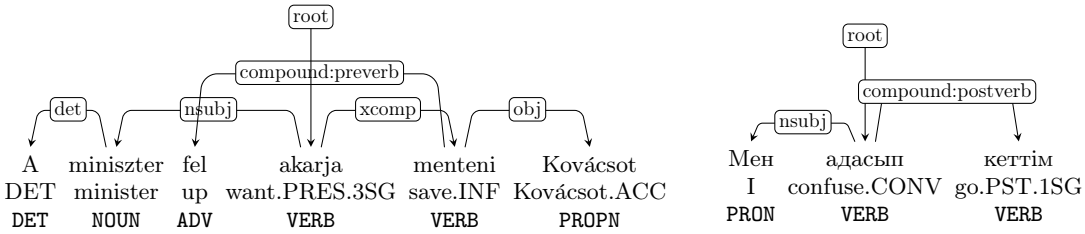
Firstly, distinguishing postverbial constructions from superficially similar multiverb constructions can be challenging. This often involves determining whether the second verb functions as a lexical verb or an auxiliary. For instance, as illustrated in (6), verbs like `yubor` may carry the lexical meaning 'to send' or modify an actional content, as in 'to do immediately and easily.' So, the phrase `ber-ib yubor` (give-CONV send) can be interpreted either as the lexical action 'to send,' i.e., 'to send through someone,' or as an actional modification of `ber` ('to give'), meaning 'to give immediately.'

Secondly, combinations of postverbial constructions can further complicate analysis. For example, in Uzbek, the phrase `yoz-ib ber-a qol` (write-CONV give-CONV remain) combines `yoz-ib ber` ('to write for someone') with `qol` ('to remain') to mean 'to start writing for someone' (Kononov, 1960, 268).

Such ambiguities significantly complicate both analysis and annotation. For instance, in the Kyrgyz-KTMU treebank, two postverbial constructions within the same sentence are analyzed differently. As shown in Figure 3, `оку-п чык` (read-CONV emerge) 'to read thoroughly' is annotated with the `advcl` relation, whereas `айт-ып бер` (tell-CONV give) 'to tell somebody' is annotated with the `ccomp` relation. Similar inconsistencies are also observed in several Turkish treebanks.

Approach	Treebank	Head Type	Cross-linguistic Applicability	Compliance with UD Guidelines	Frequency in Treebanks
advcl	Turkish-BOUN Turkish-Penn Turkish-Kenet Turkish-GB Kyrgyz-KTMU	postverb	no	no	high
xcomp/ccomp	Uyghur-UDT Kazakh-KTB	postverb	no	no	medium
aux	Azerbaijani-TueCL Kyrgyz-TueCL Tatar-NMCTT	converb	yes	yes	low
compound	Uzbek-UT	postverb	yes	yes	low

Table 2: Summary of annotation approaches for Turkic postverb constructions, detailing head type, cross-lingual applicability, compliance with UD guidelines, and the frequency of each approach across treebanks.



'The minister wants to exonerate Kovács.'

(a) Annotation for a Hungarian preverbal construction.

'I got lost.'

(b) Annotation for a Kazakh postverbial construction.

Figure 7: Possible annotation of a postverbial construction using `compound:postverb`, analogous to `compound:preverb`.

4 Discussion

The summary of the approaches described in Section 3 with their advantages and disadvantages is given in Table 2.

Tagging the converb as an adverbial clause or clausal complement while assigning the postverb as the head misrepresents the tight syntactic and semantic integration of Turkic postverb constructions. Although these two methods highlight that the converb conveys the primary lexical meaning, and are relatively common among Turkic treebanks, they do not fully adhere to UD guidelines or cross-linguistic annotation practices.

Tagging the converb as the head and the postverb as `aux` can be a reasonable approach in some contexts. In many languages, auxiliaries are desemanticized elements that support the main verb. This pattern can apply to Turkic postverbs when they primarily serve grammatical functions. However, in some Turkic languages, they might retain sufficient lexical meaning or syntactic independence to argue against classifying them as auxiliaries. For instance, if postverbs retain a significant degree of lexical meaning, a different relation such as `compound:lvc` or `compound:svc` might be more accurate.

Each of these methods has its strengths and limitations. A potential alternative could be to introduce a new language-specific subtype relation, such as `compound:postverb`, mirroring the logic behind the `compound:preverb` relation used in the Hungarian treebank (Vincze et al., 2010). This approach would avoid the misapplication of generic relations like `compound`. Figure 7 illustrates the proposed `compound:postverb` relation alongside the Hungarian example annotated with `compound:preverb`.

5 Concluding Remarks

We agree that the best approach to annotating Turkic postverbial constructions depends on

the specific properties of the language and the constraints of the annotation framework. Based on our analysis, the `compound` approach seems to be the most suitable, but we propose a dedicated subtype, `compound:postverb`, to balance semantic accuracy, syntactic clarity, and cross-linguistic comparability within the UD framework. We emphasize the importance of collaborative discussions among UD contributors, including cross-lingual and cross-treebank exchanges, to ensure robust annotation guidelines. In the future, we plan to organize a shared task within a UD Working Group to identify the optimal solution and validate the proposed annotation approach. Consistent tagging across languages and treebanks will strengthen the universality of UD, support typological linguistic studies, and foster cross-lingual applications in natural language processing (NLP).

References

Arofat Akhundjanova. 2024. UD Uzbek UT. https://github.com/UniversalDependencies/UD_Uzbek-UT.

Sevgi Ağcagül. 2004. Grammaticalization of turkic postverbial construction. Orientalia Suecana, 53:5–14.

Ibrahim Benli. 2020. UD Kyrgyz KTMU. https://github.com/UniversalDependencies/UD_Kyrgyz-KTMU.

Geert Booij and Ans Van Kemenade. 2003. Preverbs: an introduction, pages 1–11. Springer Netherlands, Dordrecht.

Neslihan Cesur, Aslı Kuzgun, Olcay Taner Yıldız, Büşra Marşan, Neslihan Kara, Bilge Nas Arıcan, Merve Özçelik, and Deniz Baran Aslan. 2022. UD Turkish Penn. https://github.com/UniversalDependencies/UD_Turkish-Penn.

Bermet Chontaeva and Çağrı Çöltekin. 2024. UD Kyrgyz TueCL. https://github.com/UniversalDependencies/UD_Kyrgyz-TueCL.

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.

Soudabeh Eslami and Çağrı Çöltekin. 2024. UD Azerbaijani TueCL. https://github.com/UniversalDependencies/UD_Azerbaijani-TueCL.

Lars Johanson. 2021. The Structure of Turkic. In Lars Johanson and Éva Á. Csató, editors, The Turkic languages, pages 26–59. Routledge.

A.N. Kononov. 1960. Grammatika sovremennogo uzbekskogo literaturnogo jazyka [Grammar of the Modern Uzbek Literary Language]. Moskva: Izdatel'stvo Akademii Nauk SSSR.

Aslı Kuzgun, Neslihan Cesur, Olcay Taner Yıldız, Oğuzhan Kuyrukçu, Arife Betül Yenice, Bilge Nas Arıcan, and Ezgi Sanıyar. 2022. UD Turkish Kenet. https://github.com/UniversalDependencies/UD_Turkish-Kenet.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 4034–4043, Marseille, France. European Language Resources Association.

Chihiro Taguchi. 2022. Consistent grammatical annotation of Turkic languages for more universal Universal Dependencies. In 29th International Conference on Head-Driven Phrase Structure Grammar.

Chihiro Taguchi, Sei Iwata, and Taro Watanabe. 2022. Universal Dependencies treebank for Tatar: Incorporating intra-word code-switching information. In Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference, pages 95–104, Marseille, France. European Language Resources Association.

Francis Tyers, Jonathan Washington, Çağrı Çöltekin, and Aibek Makazhanov. 2017. An assessment of universal dependency annotation guidelines for turkic languages. In Proceedings of the 5th International Conference on Turkic Languages Processing (TurkLang 2017), pages 276–297.

Francis M. Tyers and Jonathan N. Washington. 2015. Towards a free/open-source universal-dependency treebank for kazakh. In 3rd International Conference on Turkic Languages Processing, (TurkLang 2015), pages 276–289.

Utku Türk, Furkan Atmaca, Şaziye Betül Özateş, Gözde Berk, Seyyit Talha Bedir, Abdullatif Köksal, Balkız Öztürk Başaran, Tunga Güngör, and Arzucan Özgür. 2021. Resources for turkish dependency parsing: Introducing the boun treebank and the boat annotation tool. Preprint, arXiv:2002.10416.

Veronika Vincze, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. 2010. Hungarian dependency treebank. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA).

Jonathan Washington, Çağrı Çöltekin, Furkan Akkurt, Bermet Chontaeva, Soudabeh Eslami, Gulnura Jumalieva, Aida Kasieva, Aslı Kuzgun, Büşra Marşan, and Chihiro Taguchi. 2024. Strategies for the annotation of pronominalised locatives in Turkic Universal Dependency treebanks. In Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024, pages 207–219, Torino, Italia. ELRA and ICCL.

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.