Consistent Grammatical Annotation of Turkic Languages for more universal Universal Dependencies

Chihiro Taguchi
University of Edinburgh / Nara Institute of Science and Technology
Edinburgh, United Kingdom / Nara, Japan

1 Introduction

Recent advance in natural language processing (NLP) has shed light on its universal application to any languages, in particular languages with less resources. Universal Dependencies (UD) (Nivre et al., 2020) is one of such cross-linguistic projects that aims to accumulate annotated linguistic data with the rules that can ideally be universally applied to and are consistent in any languages. For cross-linguistic applications and research in NLP and linguistics, it is of paramount importance to use annotated linguistic data with unified annotation rules. However, in practice, current UD treebanks contain a number of discords in terms of annotation rules, even in the same language. This disagreement in the annotation style hinders simultaneous use of multiple treebanks because it is likely to cause confusion in, for instance, supervised tasks and transfer learning.

Given this situation, this study reports current annotation conventions of tokenization, part-of-speech (POS), and morphological tagging and their conflicts in the UD treebanks of Turkic languages as a case study. For each issue, this study provides possible solutions that are linguistically justifiable for unified annotation rules that commonly apply to Turkic languages.

Annotation consistency assessment in UD has been quantitatively investigated in de Marneffe et al. (2017) for UD English, French, and Finnish; however, it solely evaluates dependency head and relation, and other aspects of UD annotation are trimmed away. Although Tyers et al. (2017) assesses the guidelines for five treebanks of three Turkic languages in details, the number of treebanks and Turkic languages in UD has doubled since its publication. Therefore, it is worthwhile to investigate the (dis)agreements in the annotation across Turkic UD once again.

1.1 Turkic Languages

The Turkic language family is a genetically related group of languages distributed across Eurasia, from Turkish in the west to Sakha (Yakut) and Uyghur in East Asia, as illustrated in Figure 1. The hypothetical Proto-Turkic, which modern Turkic languages commonly derived from, is estimated to have emerged around 4500–4000 BCE (Johanson, 2021). Modern Turkic languages are roughly divided into four groups and two isolated languages: Southwestern (Turkish, Azeri, Turkmen, etc.), Northwestern (Tatar, Kazakh, Kyrgyz, etc.), Southeastern (Uyghur, Uzbek, etc.), Northeastern (Sakha, Tuvan, etc.), Chuvash and Khalaj. Despite its broad geographical distribution and long history, Turkic languages are known to share typological, phonological, morphological, syntactic, and lexical characteristics. Most of them commonly have SOV word order, vowel harmony, inflection and derivation by agglutinative suffixation, and shared lexicon. For this reason, building multilingual treebanks for Turkic languages may enable not only applications in NLP but also quantitative studies in linguistics. However, as the following sections show, there is discrepancies in annotation among the Turkic treebanks.

1.2 Turkic UD Treebanks Discussed

In this paper, we will focus on the UD treebanks of Turkic languages published as of version 2.10 (updated in May 2022). The list of treebanks are shown in Table 1. There are only seven Turkic languages covered in current UD. Turkish is by far the most high-resourced language among the Turkic UD treebanks with more than 600K tokens in total. Turkish–German SAGT is a spoken code-switching treebank of Turkish mixed with German used by the Turkish community in Germany. Old Turkic[1], categorized in the Northeastern Turkic group, is an extinct language used around the 10th centuries around current Mongolia and Xinjiang, and is the earliest Turkic language attested in history.

2 Issues

Each UD treebank is supposed to contain at least the following grammatical information: token (`WORD`), lemma (`LEMMA`), part-of-speech (`UPOS`), morphological features (`FEATS`), head of the word (`HEAD`), and dependency relation (`DEPREL`). In this study, we will particularly delve into differences in tokenization, `UPOS`, and `FEATS`, because their annotation tend to vary among the treebanks and also they play significant roles in training for tagging tasks.

[1]Though the Tonqq treebank uses *Old Turkish* as the language name, this study refers to it as *Old Turkic* because the latter is a more common convention in Turkic linguistics; also, the former can falsely remind us of a direct ancestor of the modern Turkish language, which is comparatively distant from Old Turkic.

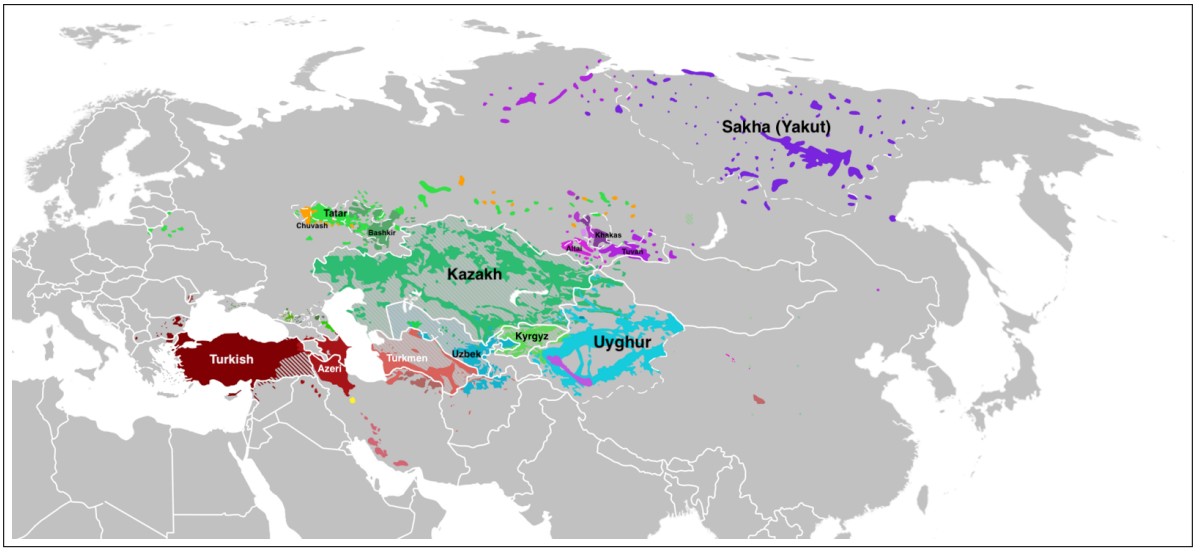

Figure 1: Distribution of Turkic languages.

2.1 Tokenization

Among Turkic UD treebanks, there are two issues in tokenization that has yet to be agreed upon: whether to tokenize a suffix *-ki* (e.g., Turkish GB) as a separate token and whether to tokenize all suffixes (morphemes) as different tokens (e.g., Old Turkic Tonqq).

Adjectivizer *-ki*.

Many Turkic languages have an adjectivizer suffix *-ki* (or phonologically corresponding forms). It is used for a locative noun to modify a following noun, as exemplified in an example in Turkish (1). However, because it is possible to treat the adjectivized form as a noun, it theoretically allows a recursive use of the suffix (2).

(1) *Berlin-de-ki* (kişi)
Berlin-LOC-KI person
'(The person) in Berlin'

(2) [[*Berlin-de-ki*]-*ler-de-ki*]-ler-...
Berlin-LOC-KI-PL-LOC-KI-PL-...
'... those that are at those in Berlin's'

Considering that UD's morphological features (FEAT) do not allow for hierarchical annotation, Turkic treebanks such as GB and SAGT use a trick to handle this recursion. These treebanks retokenize *-ki* as a separate token and gives it an ADP (adposition) tag, as shown in Table 2. While this policy can avoid the recursion problem, there are still remaining issues. First, it does not seem to be possible to justify treating this particular suffix *-ki* as a separate token but not others. Second, it is questionable to tag *ki* with ADP even though its function is an adjectivalizer. This function seems rather suitable to be annotated as amod (adjectival modification) in DEPREL as in Table 3. The limitation in the expressiveness of morphology is caused by UD's architecture that annotates FEAT only linearly. It can be fixed by allowing for hierarchical morphological annotation that has been attempted in Unimorph 4.0 (Batsuren et al., 2022).

Old Turkic.

The tokenization policy in the Old Turkic Tonqq treebank is fundamentally different from other Turkic UD treebanks. Because Old Turkic texts written in the Old Turkic script do not explicitly show word segmentation by spacing unlike other modern Turkic orthographies, conventional approaches of tokenization do not apply to Old Turkic. Instead, the segmentation policy taken in the Tonqq treebank is to segment every morpheme (Derin and Harada, 2021). This is the same tactic used in UD treebanks of Japanese, which also does not have orthographical system for delimiting words. For example, a cognate inflected finite verb сагындым 'I had a longing (for)' in Tatar or *saqntm* 'I thought over' in Old Turkic are annotated as (4) and (5), respectively. It is controversial as to which tokenization is linguistically more suitable; however, in practice, there are several weaknesses in the morpheme-splitting tokenization. First, other Turkic languages in UD do not split by morpheme boundaries. Therefore, splitting by morphemes can cause inconsistency across the Turkic UD treebanks. Second, this policy makes the difference between an independent word and a bound morpheme ambiguous. Third, as one can see by comparing Tables 4 and 5, the latter virtually tells us nothing about the morphological features added by the suffixes (tagged as AUX in the Table).

Group	Language	Treebank	Latest	Tokens	Annotation	Source
SW	Turkish	Kenet (Kuzgun et al., 2022b)	v2.10	178K	Manual	Dictionary
		Penn (Cesur et al., 2022)	v2.10	87K	Manual	Penn Treebank
		Tourism (Kuzgun et al., 2022a)	v2.10	92K	Manual	Reviews
		Atis (Köse and Yıldız, 2022)	v2.10	45K	Semi-auto	ATIS
		FrameNet (Marşan et al., 2021)	v2.9	19K	Manual	FrameNet
		GB (Çöltekin, 2020)	v2.7	17K	Manual	Grammar book
		IMST (Çöltekin et al., 2021)	v2.8	57K	Semi-auto	IMST Treebank
		BOUN (Türk et al., 2020)	v2.8	122K	Manual	Miscellaneous
		PUD (Uszkoreit et al., 2021)	v2.8	16K	Semi-auto	PUD
	Turkish-German	SAGT (Çetinoğlu, 2016)	v2.10	37K	Manual	Spoken
SE	Uyghur	UDT (Eli et al., 2016)	v2.8	40K	Manual	Books
NW	Kazakh	KTB (Makazhanov et al., 2015)	v2.10	10K	Manual	Miscellaneous
	Tatar	NMCTT (Taguchi et al., 2022)	v2.10	1K	Manual	News
NE	Sakha (Yakut)	YKTDT (Merzhevich and Gerardi, 2021)	v2.10	495	Manual	Miscellaneous
	Old Turkic	Tonqq (Derin and Harada, 2021)	v2.10	158	Manual	Inscriptions

Table 1: Details of the Turkic treebanks available as of UD v2.10. Semi-auto in the Annotation column means that the annotation was done by combining an automatic tagging process with manual annotation.

ID	FORM	LEMMA	UPOS	DEPREL
1-2	Berlin'deki	_	_	_
1	Berlin'de	Berlin	PROPN	nmod
2	ki	ki	ADP	case

Table 2: Tokenization and tags of *ki* as a separate token.

ID	FORM	LEMMA	UPOS	DEPREL
1	Berlin'deki	Berlin	PROPN	amod

Table 3: Tokenization and tags of *ki* as morphologically suffixed to the stem.

2.2 Part-of-Speech Tagging

Particle.
In UD, the particle tag PART is rather not recommended to use; the general guideline notes that it should be used when no other tags are suitable[2]. Turkic treebanks that use PART are BOUN (Turkish), SAGT (Turkish–German), UDT (Uyghur), and KTB (Kazakh) as listed in Table 6.

The general guideline of Turkish UD treebanks defines that they should "use the POS tag PART for the word *değil* 'not' when used to negate a non-predicate word", and "[i]f değil modifies a predicate, it is marked as VERB since it functions as a copula and carries other verbal inflections as well".[3]d However, BOUN, as well as the treebanks not listed in Table 6, indicate that this guideline is not strictly followed by them. As for SAGT, it consistently marks *değil* with PART regardless of the usage specified above.

What falls into PART in Turkic languages' treebanks seems inconsistent and incompatible with other Turkic treebanks. For example, KTB marks *ma* (yes-no question word) and its harmonized counterparts (*ba*, *me*, *be*) with PART; however, their counterparts in other Turkic languages, for instance Turkish, are generally annotated as AUX. Following UD's general policy that tries to avoid the use of PART, it is favorable to use other alternatives such as AUX and ADV where applicable. In fact, by comparing with other Turkic treebanks that avoid PART, many PART words in Uyghur-UDT and Kazakh-KTB can be substituted with AUX or ADV. Since the status of these closed POS categories is controversial even in linguistics, what we need for UD is a unified policy for tagging them. This unification process is possible through cross-treebank and cross-linguistic comparison and consultation with linguists.

Converb: VERB or ADV?
Converb, also called adverbial participle, is a non-finite form of a verb and it adverbially modifies other (typically main) predicates (Haspelmath, 1995). In Turkic languages, it is common to use this form to express consecutive or simultaneous processes. In Table 7, we can see that the same converb form is annotated either as VERB or ADV. However, since converb is a productively inflected form of a verb, it is plausible to tag it with VERB rather than ADV.

Locative adjectivizer: NOUN/PROPN or ADJ?
A similar problem is also found in the treatment of *-ki* following a locative suffix, as we have seen in Section

[2]https://universaldependencies.org/u/pos/PART.html
[3]https://universaldependencies.org/tr/pos/PART.html

ID	FORM	LEMMA	UPOS	FORM
1	сагындым	сагын	VERB	Number=Sing|Person=1|Tense=Past|VerbForm=Fin

Table 4: Tokenization of сагындым (Tatar; transliterated as *sa*ɣ*ındım*) split by space.

ID	FORM	LEMMA	UPOS	FORM
1	saqn	_	VERB	_
2	t	_	AUX	_
3	m	_	AUX	_

Table 5: Tokenization of *saqntm* (Old Turkic; reconstructed as *saqïntïm*) in Tonqq's policy. The tokens are transliterated.

Treebank	PART words
BOUN	*ki* (that)
SAGT	*değil* (not), *nicht* (not)
UDT	*de* (also), *qëni* (well), *belkim* (maybe), *epsus* (pity), *mana* (here), *emesmu* (not), *goya* (as if), *ëhtimal* (probably), etc.
KTB	*ma* (yes-no question), *aw* (emphatic), *šïɣar* (probably), *qoy* (emphatic), etc.

Table 6: List of treebanks that use PART and their PART words. Uyghur and Kazakh words are transliterated into Latin for convenience.

Treebank	UPOS	FEAT
Kenet	ADV	_
Penn	ADV	_
Tourism	ADV	_
Atis	ADV	_
GB	VERB	VerbForm=Conv
FrameNet	ADV	_
BOUN	VERB	_
PUD	ADV	_
IMST	VERB	VerbForm=Conv
SAGT	VERB	VerbForm=Conv
KTB	VERB	VerbForm=Conv
NMCTT	VERB	VerbForm=Conv
UDT	VERB	VerbForm=Conv
YKTDT	NA	NA

Table 7: UPOS and FEAT annotation for converbs. An underscore means no annotation given to the form, and NA means converb is unattested in the corpus.

2.1. Though *Berlin-de-ki* adjectivally modifies a noun at the level of dependency relation, it should be tagged as PROPN (i.e., the lemma's POS), because it is a productively inflected from a nominal. However, there are several Turkish UD treebanks that tag these as ADJ, such as Kenet and Penn.

2.3 Morphological Tagging

Bare noun.
In Turkic languages, the default unmarked noun form is nominative singular (Case=Nom|Number=Sing), and is agreed by 3rd person (Person=3) in verbal morphology. How each treebank annotates the bare form of a noun greatly differs among each other as demonstrated in Table 8. Given the fact that the bare form is morphologically unmarked, we may well omit any morphological annotation on it. Nevertheless, with the interpretation that the nominative and singular forms are a part of the nominal paradigm, we can also exhaustively write them out, thereby improving the consistency with other declined noun forms. Person=3, on the other hand, is not a suitable morphological feature for a bare noun, because the noun in Turkic languages does not morphologically change with respect to person[4] unlike verbs and personal pronouns. In sum, the annotation of a bare noun can be minimally blank, but it seems better to annotate it as Case=Nom|Number=Sing.

Converb.
Returning to the distribution of the annotation for converbs in Table 7, we can find another problem in the FEAT column. That is, many treebanks of Turkish do not annotate any morphological information in the FEAT column. Since UD readily prepares VerbForm=Conv tag for converbs[5], it is favorable to follow the general guideline.

Furthermore, among the treebanks that mark VerbForm=Conv, some attempt to distinguish converb forms by further specifying the morphological feature. For example, in Kazakh-KTB, it prepares Aspect=Imp|VerbForm=Conv and Aspect=Perf|VerbForm=Conv to distinguish two

[4]Except for possessor, which is marked by another feature Person[psor]=.

[5]https://universaldependencies.org/tr/feat/VerbForm.html

Treebank	Case=Nom	Number=Sing	Person=3
Kenet	Y	Y	Y
Penn	Y	Y	Y
Tourism	Y	Y	Y
Atis	Y	Y	Y
GB	Y	Y	N
FrameNet	Y	Y	Y
BOUN	Y	Y	Y
PUD	Y	N	Y
IMST	Y	Y	Y
SAGT	Y	Y	N
UDT	Y	N	N
KTB	Y	N	N
NMCTT	Y	Y	N
YKTDT	Y	N	N
tonqq	N	N	N

Table 8: Comparison of annotation for a bare noun.

forms of converb that function differently in terms of aspect. This is another issue to be discussed in more depth in treebanks of languages that have a converb.

3 Concluding Remarks

This study showed that the current Turkic treebanks contain disagreements and inconsistencies in the rules of tokenization, UPOS, and FEATS. To these issues, we also showed favorable solutions based on linguistic evidence and comparisons with other treebanks. The conflicts of annotation in words that are cognate or function in the same manner should be avoided as much as possible. Therefore, we call on UD contributors for cross-lingual and cross-treebank discussions. Also, consultation with linguists specialized in the language or related languages is indispensable for defining justifiable and plausible rules for consistent annotation. We hope that coherent tagging across languages and treebanks in UD will further enhance the universality of UD, thereby enabling UD-based typological studies in linguistics and encouraging cross-treebank and cross-lingual applications in NLP.

References

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold KieraÅ›, GÃ¡bor 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, EsaÃº Zumaeta Rojas, Didier LÃ³pez Francis, Arturo Oncevay, Juan LÃ³pez 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, RitvÃ¡n KarahÃ³Ç§a, 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 Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

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.

Özlem Çetinoğlu. 2016. A Turkish-German code-switching corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4215–4220, Portorož, Slovenia. European Language Resources Association (ELRA).

Marie-Catherine de Marneffe, Matias Grioni, Jenna Kanerva, and Filip Ginter. 2017. Assessing the annotation consistency of the Universal Dependencies corpora. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 108–115, Pisa,Italy. Linköping University Electronic Press.

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.

Martin Haspelmath. 1995. The converb as a cross-linguistically valid category. *Converbs in Cross-linguistic Perspective*.

Lars Johanson. 2021. *Turkic*. Cambridge Language Surveys. Cambridge University Press.

Aslı Kuzgun, Neslihan Cesur, Olcay Taner Yıldız, Oğuzhan Kuyrukçu, Büşra Marşan, Bilge Nas Arıcan, Neslihan Kara, Deniz Baran Aslan, Ezgi Sanıyar, and Cengiz Asmazoğlu. 2022a. UD Turkish Tourism. https://github.com/UniversalDependencies/UD_Turkish-Tourism.

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

Mehmet Köse and Olcay Taner Yıldız. 2022. UD Turkish Atis. https://github.com/UniversalDependencies/UD_Turkish-Atis.

Aibek Makazhanov, Aitolkyn Sultangazina, Olzhas Makhambetov, and Zhandos Yessenbayev. 2015. 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, Neslihan Kara, Merve Özçelik, Bilge Nas Arıcan, Neslihan Cesur, Aslı Kuzgun, Ezgi Sanıyar, Oğuzhan Kuyrukçu, and Olcay Taner Yildiz. 2021. Building the Turkish FrameNet. In *Proceedings of the 11th Global Wordnet Conference*, pages 118–125, University of South Africa (UNISA). Global Wordnet Association.

Tatiana Merzhevich and Fabrício Ferraz Gerardi. 2021. UD Yakut YKTDT. https://github.com/UniversalDependencies/UD_Yakut-YKTDT.

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 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

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 (EURALI)*, Marseille, France.

Francis M. 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)*.

Utku Türk, Furkan Atmaca, ̧Saziye Betül Özateş, Gözde Berk, Seyyit Talha Bedir, Abdullatif Köksal, Balkız Öztürk Ba ̧saran, Tunga Güngör, and Arzucan Özgür. 2020. Resources for Turkish Dependency Parsing: Introducing the BOUN Treebank and the BoAT Annotation Tool.

Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Slav Petrov, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Savas Cetin, Martin Popel, Daniel Zeman, Francis Tyers, Çağrı Çöltekin, Utku Türk, Furkan Atmaca, Şaziye Betül Özateş, Abdullatif Köksal, Balkız Öztürk Başaran, Tunga Güngör, and Arzucan Özgür. 2021. UD Turkish PUD. https://github.com/UniversalDependencies/UD_Turkish-PUD.

Çağrı Çöltekin. 2020. UD Turkish GB. https://github.com/UniversalDependencies/UD_Turkish-GB.

Çağrı Çöltekin, Gülşen Cebiroğlu Eryiğit, Memduh Gökırmak, Hüner Kaşıkara, Umut Sulubacak, and Francis Tyers. 2021. UD Turkish IMST. https://github.com/UniversalDependencies/UD_Turkish-IMST.