UD-Tatar NMCTT Treebank: Issues in Annotation across Turkic UD

Chihiro Taguchi

Natural Language Processing Group
Department of Computer Science and Engineering, University of Notre Dame
https://ctaguchi.github.io

September 8, 2023

Outline

Introduction

Issues and Solutions

- Tokenization
- Part-of-speech
- Morphology

Conclusion

Introduction

This report:

- introduces UD Tatar NMCTT Treebank
- Discusses annotation disagreements across Turkic UD treebanks, with a special focus on Tatar

This report is based on my presentation at the Workshop on Computational Linguistics on East Asian languages in 2022. Some examples from treebanks are from UD v2.10; please correct me if the data used are outdated.

Turkic Languages

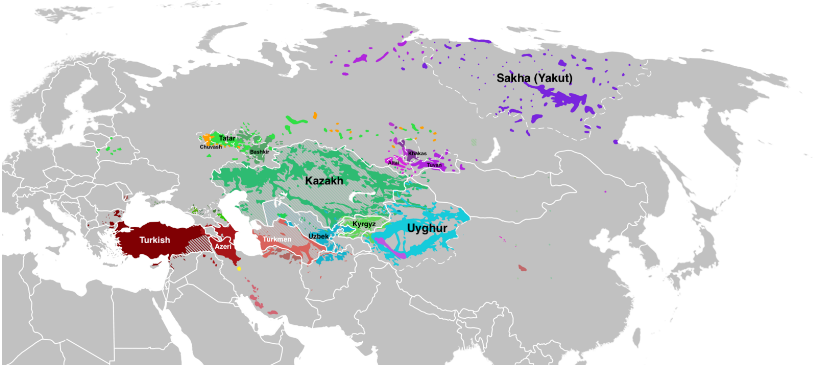

Figure: Distribution of the Turkic languages

Turkic Languages and UD

Southwestern (Oghuz)

- Turkish (Turkey; ~80 mil.)

Northwestern (Kipchak)

- Kazakh (Kazakhstan; ~18 mil.)
- Tatar (Republic of Tatarstan, Russia; ~5 mil.)
- Kyrgyz (Kyrgyzstan; ~4 mil.)

Southeastern (Karluk)

- Uyghur (Xinjiang, China; ~10 mil.)

Northeastern (Siberian)

- Sakha (Republic of Sakha, Russia; ~0.5 mil.)
- Old Turkic (Current Mongolia; Extinct)

Turkic treebanks in UD

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)	v2.11	57K	Semi-auto	IMST Treebank
		BOUN (Türk et al., 2020)	v2.11	122K	Manual	Miscellaneous
		PUD (Uszkoreit et al., 2021)	v2.12	16K	Semi-auto	PUD
	Turkish-German	SAGT (Çetinoğlu and Çöltekin, 2022)	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
	Kyrgyz	KTMU (Benli, 2023)	v2.12	7K	Manual	News etc.
	Tatar	NMCTT (Taguchi et al., 2022)	v2.11	1K	Manual	News
NE	Sakha (Yakut)	YKTDT (Merzhevich and Gerardi, 2021)	v2.10	495	Manual	Miscellaneous
	Old Turkish (Old Turkic)	Tonqq (Derin and Harada, 2021)	v2.10	158	Manual	Inscriptions

Table: 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.

UD Tatar NMCTT

- The first and only treebank for Tatar
- NMCTT: NAIST Multilingual Corpus Tatar (funded by Nara Institute of Science and Technology)
- Online news text (tatar.inform) with the condition that the source link to the news page is shown
- 148 sentences, 2280 tokens
- Marks code-switching and its boundaries (will be incorporated to the SAGT style)

ID	FORM	LEMMA	UPOS	FEATS	HEAD	DEPREL	MISC
1	Татарстанда	Татарстан	PROPN	Case=Loc|Number=Sing	5	obl	LangID=TT
2	коронавирустан	коронавирус	NOUN	Case=Abl|Number=Sing	4	nmod	CSPoint=коронавирус§тан|LangID=MIXED[TT§RU]
3	беренче	беренче	ADJ	_	4	amod	LangID=TT
4	прививканы	прививка	NOUN	Case=Acc|Number=Sing	5	obj	CSPoint=прививкас§ны|LangID=MIXED[TT§RU]
5	ясатырга	яса	VERB	VerbForm=Inf|Voice=Cau	0	root	LangID=TT
6	мөмкин	мөмкин	AUX	_	5	aux	LangID=TT|SpaceAfter=No
7	.	.	PUNCT	_	5	punct	LangID=OTHER

Table: An example of UD Tatar. The sentence is transcribed as *Tatarstanda koronavirustan berençe privivkanı yasatırğa mömkin.*

Code-switching annotation

Frequent code-switching with Russian: NMCTT marks the language code(s) and the code-switching boundary in the MISC column.

- LangID: Language ID (will be changed to CSID and lang from the next version to be unified with SAGT)
- CSPoint: Code-switching point. § marks the code-switching boundary.
- If LangID=MIXED, the segment-level LangIDs are shown in brackets.
  - коронавирус§тан: LangID=MIXED[RU§TT] "from coronavirus"
  - гомум§техник: LangID=MIXED[TT§RU] "polytechnic"

ID	FORM	LEMMA	UPOS	FEATS	HEAD	DEPREL	MISC
1	Татарстанда	Татарстан	PROPN	Case=Loc|Number=Sing	5	obl	LangID=TT
2	коронавирустан	коронавирус	NOUN	Case=Abl|Number=Sing	4	nmod	CSPoint=коронавирус§тан|LangID=MIXED[TT§RU]
3	беренче	беренче	ADJ	_	4	amod	LangID=TT
4	прививканы	прививка	NOUN	Case=Acc|Number=Sing	5	obj	CSPoint=прививкас§ны|LangID=MIXED[TT§RU]
5	ясатырга	яса	VERB	VerbForm=Inf|Voice=Cau	0	root	LangID=TT
6	мөмкин	мөмкин	AUX	_	5	aux	LangID=TT|SpaceAfter=No
7	.	.	PUNCT	_	5	punct	LangID=OTHER

Table: An example of UD Tatar. The sentence is transcribed as *Tatarstanda koronavirustan berençe privivkanı yasatırğa mömkin.*

Issues in Turkic UD

Issues I faced during my annotation of Tatar UD

Tokenization

- Locative adjectivizer *-ki*: separate token?
- Tokenization by word vs. by morpheme

Part-of-speech

- Usage of particles `PART`
- Converbs: `VERB` or `ADV`?
- Adjectivized locative nouns: `ADJ` or `NOUN`?

Morphology

- Bare noun: `Case=Nom|Number=Sing` or nothing?
- Converb: `VerbForm=Conv`

Tokenization: Turkic locative adjectivizer *-ki*

-ki

- Shared across many Turkic languages
- Used for a locative noun to modify the next noun (1)
- Recursion (2)
- UD's flat (non-nested) annotation of morphological features cannot handle this

(1) Turkish
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'

Tokenization: Turkic locative adjectivizer *-ki*

Solutions

1. *-ki* as an independent token with the `case` relation (e.g., GB and SAGT; Table)
2. *-ki* as an independent token with the `dep:der` relation (ATIS)
3. Introduce hierarchical morphological annotation (cf. UniMorph 4.0 (Batsuren et al., 2022))

In UD-Tatar NMCTT

-ki is not tokenized separately (also in Kyrgyz KTMU)

- This still does not solve the nested features (Turkish *ev-ler-de-ki-ler-de-...* or Tatar *öy-lär-dä-ge-lär-dä-...*)
- Any suggestions?

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

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

ID	FORM	LEMMA	UPOS	FEAT	DEPREL
1	Берлиндагы	Берлин	PROPN	Case=Loc	amod

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

Tokenization: word vs. morpheme

Word boundary in orthographies

- Modern Turkic languages: syntactic word boundaries are split by spaces in the orthography (+ some exceptions)
- Old Turkic: word boundaries are not necessarily split in the original Old Turkic inscription text

Inscription:	𐰘𐱅𐰃𐰘𐰇𐰔𐰉𐰆𐰡𐰃
Transcription	j^2 t^2 i j^2 ü z b^1 o lt i
Reconstruction[1]:	*jeti jüz boltï*

[1] Reconstruction by the Language Committee of Ministry of Culture and Information of the Republic of Kazakhstan

Tokenization: word vs. morpheme

Tokenization in the Old Turkic Tonqq Treebank

- Old Turkic Tonqq treebank tokenizes every morpheme.

(3) 𐰘𐱅𐰃𐰘𐰇𐰔𐰉𐰆𐱃𐰃

jeti	*jüz*	*bol-tï*
seven	hundred	be-PST

‘(They were) seven hundred’

FORM	LEMMA	UPOS	FEAT
𐰘𐱅𐰃	_	NUM	_
𐰘𐰇𐰔	_	NUM	_
𐰉𐰆𐰞	_	VERB	_
𐱄𐰃	_	AUX	_

Table: Tonqq-style tokenization of (3).

FORM	lemma	UPOS	FEAT
𐰘𐱅𐰃	𐰘𐱅𐰃	NUM	_
𐰘𐰇𐰔	𐰘𐰇𐰔	NUM	_
𐰉𐰆𐱃𐰃	𐰉𐰆𐰞	VERB	Tense=Past

Table: Conventional tokenization of (3).

Tokenization: word vs. morpheme

Which is suitable in UD?

- Controversial in linguistics
- This study's position: word-splitting tokenization

Why not morpheme-splitting tokenization

- Inconsistent: Other Turkic languages in UD do not split by morphemes
- Ambiguous: The difference between independent words and bound morphemes is less clear
- Less informative: The current morpheme-split method does not tell us anything about morphological features

Part-of-speech: Particle `PART`

In general, the `PART` tag should be used restrictively and only when no other tag is possible. — UD Guideline

Use of `PART` in Turkic UD

- Not unified, as in the Table
- UD Turkish's guideline defines that only *değil* negating non-predicate word is `PART`
- Actual usage varies in every treebank

Treebank	`PART` words
BOUN	*ki* (that), *çok* (much), and other 61 words
SAGT	*değil* and other 7 German words
UDT	*de* (also), *qëni* (well), *belkim* (maybe), *epsus* (pity), and other 18 words
KTB	*ma* (yes-no question particle), *šïɣar* (probably), and other 7 words
KTMU	None
NMCTT	None

Table: Usage of `PART` and their `PART` words. Uyghur and Kazakh words are transliterated for convenience.

Part-of-speech: Particle `PART`

> *In general, the `PART` tag should be used restrictively and only when no other tag is possible.* — UD Guideline

Solution

- The status of these closed-class POS categories is controversial in linguistics
- For UD, having a unified policy is better than being inconsistent
- Following the guideline, try not to use `PART`
- Instead, use `AUX`, `ADV`, or other feasible tags

Part-of-speech: Converb

Converb

- Verb form modifying other predicates adverbially (Haspelmath, 1995)
- Also called “adverbial participle”
- Disagreement: VERB or ADV?

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
KTMU	VERB	VerbForm=Conv
NMCTT	VERB	VerbForm=Conv
UDT	VERB	VerbForm=Conv
YKTDT	NA	NA

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

Part-of-speech: Converb

Issue

- VERB or ADV?

Solution

- Converbs (at least in Turkic) are VERB
- Turkic converbs are productively inflected
- In UD Tatar, the four converb forms are distinguished in FEATS
    - *-Ip*: `VerbForm=Conv`
    - *-GAnçI*: `Aspect=Imp|VerbForm=Conv`
    - *-A*: `Aspect=Prog|VerbForm=Conv`
    - *-GAç*: `Aspect=Perf|VerbForm=Conv`
- Remaining issue: Aspect conflict in grammaticalized constructions
    - Tatar: *jaz-ğal-ıy başla-dı* “S/he started to write many times”
    - `Aspect=Iter` (*-GAlA*) and `Aspect=Prog` (-A)?

Part-of-speech: Adjectivized locative noun

-*ki* again: NOUN or ADJ?

- Generally, -*ki* with locative is productively derived from noun
- Therefore, it should be tagged as NOUN (or PROPN)

Treebank	UPOS	FEAT
Kenet	ADJ	amod
Penn	ADJ	amod
Tourism	ADJ	nmod
Atis	ADJ	amod
GB	NOUN + ADP	nmod
FrameNet	NOUN + ADP	amod
BOUN	NOUN	amod
PUD	NOUN + ADP	amod
IMST	NOUN + ADP	nmod
SAGT	NOUN + ADP	nmod
KTB	NOUN	amod
KTMU	NOUN	nmod:poss, nmod
NMCTT	NOUN	amod
UDT	NOUN	amod
YKTDT	NOUN	nmod

Table: UPOS and FEAT annotation for -*ki*.

Morphology: Bare noun

Default unmarked noun form

- Nominative `Case=Nom`
- Singular `Number=Sing`
- Agrees with verbs in 3rd person `Person=3`
- Differences across Turkic treebanks

Language	Treebank	`Case=Nom`	`Number=Sing`	`Person=3`
Turkish	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
Turkish-German	SAGT	Y	Y	N
Uyghur	UDT	Y	N	N
Kazakh	KTB	Y	N	N
Kyrgyz	KTMU	Y	Y	Y
Tatar	NMCTT	Y	Y	N
Yakut	YKTDT	Y	N	N
Old Turkic	Tonqq	N	N	N

Table: Comparison of annotation for a bare noun.

Morphology: Bare noun

Solution

- Case (nominative) and number (singular) are changeable features of the nominal paradigm
- Person is not a distinctive feature of the `NOUN` paradigm (`NOUN` is always `Person=3`)
- Bare nouns should be marked as `Case=Noun|Number=Sing`

Morphology: Converb

Issue

Many of the Turkish treebanks do not tag converb forms as `VerbForm=Conv`

Solution

Converbs should be tagged as `VerbForm=Conv`

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
KTMU	VERB	VerbForm=Conv
NMCTT	VERB	VerbForm=Conv
UDT	VERB	VerbForm=Conv
YKTDT	NA	NA

Table: `UPOS` and `FEAT` annotation for converbs (-*Ip*, -*ArAk*, etc.). An underscore means no annotation given to the form in the corpus; NA means converb is unattested in the corpus.

Concluding Remarks

This study ...

- Introduced UD-Tatar NMCTT
- Showed disagreeing annotation of Turkic UD treebanks
  - Tokenization
  - POS tags
  - Morphological features
- Proposed a solution for each case

For more universal Universal Dependencies ...

- Cross-lingual and cross-treebank discussions
- Involvement of linguists

Thank you!

Contact: `ctaguchi@nd.edu`

Website: `https://ctaguchi.github.io`

References I

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. Unimorph 4.0: Universal morphology. In *Proceedings of the Language Resources and Evaluation Conference*, pages 840–855, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.89.

Ibrahim Benli. UD kyrgyz KTMU. https://github.com/UniversalDependencies/UD_Kyrgyz-KTMU/tree/master, 2023.

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

References II

Mehmet Oguz Derin and Takahiro Harada. Universal Dependencies for Old Turkish. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 129–141, Sofia, Bulgaria, December 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.udw-1.11`.

Marhaba Eli, Weinila Mushajiang, Tuergen Yibulayin, Kahaerjiang Abiderexiti, and Yan Liu. 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, December 2016. The COLING 2016 Organizing Committee. URL `https://aclanthology.org/W16-5206`.

Martin Haspelmath. The converb as a cross-linguistically valid category. *Converbs in Cross-linguistic Perspective*, 01 1995.

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

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

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

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

References III

Büşra Marşan, Neslihan Kara, Merve Özçelik, Bilge Nas Arıcan, Neslihan Cesur, Aslı Kuzgun, Ezgi Sanıyar, Oğuzhan Kuyrukçu, and Olcay Taner Yildiz. Building the Turkish FrameNet. In *Proceedings of the 11th Global Wordnet Conference*, pages 118–125, University of South Africa (UNISA), January 2021. Global Wordnet Association. URL `https://aclanthology.org/2021.gwc-1.14`.

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

Chihiro Taguchi, Sei Iwata, and Taro Watanabe. 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, 2022.

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

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. UD Turkish PUD. `https://github.com/UniversalDependencies/UD_Turkish-PUD`, 2021.

Özlem Çetinoğlu and Çağrı Çöltekin. Two languages, one treebank: building a Turkish–German code-switching treebank and its challenges. *Language Resources and Evaluation*, pages 1–35, 2022. ISSN 1574-020X. doi: 10.1007/s10579-021-09573-1.

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

References IV

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