The Turkish GB treebank

With a few notes on Turkish-German Code Switching Treebank

Çağrı Çöltekin

Department of Linguistics, University of Tübingen

September, 2023

The Turkish ‘grammar book’ treebank

- A (small) treebank consisting of grammar book examples (primarily from Göksel and Kerslake, 2005)
- 2 880 “sentences”, 17 177 tokens
- 371 words were segmented
- All sentences were manually annotated for syntax and morphology
- Each sentences is accompanied with a gloss and English translation (as in the source)
- Some entries (473) in the treebank are not full sentences
- A few entries are repeated with alternative analysis
- Annotated using UD v1, automatically converted (with some manual corrections) to UD v2
- All sentences were annotated by a single annotator

Why?

- Developing UD dependency annotation scheme for Turkish
- Wide coverage of morphosyntactic constructions with little annotation effort
- Example sentences in grammar books aim for wide coverage!

What was difficult?

(out of 2015 TLT talk)

- Tokenization: sub-word syntactic units
- Empty/null units
- Missing or incompatible morphological/lexical features
- Head direction
- Clausal or non-clausal argument distinction
- Argument-adjunct distinction

What was difficult?

(out of 2015 TLT talk)

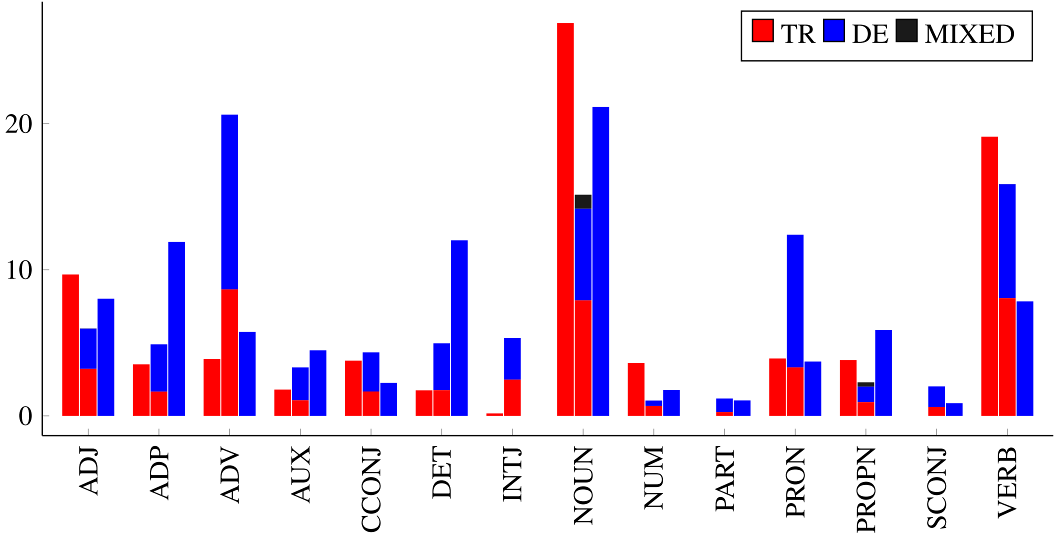
- Tokenization: sub-word syntactic units
- Empty/null units
- Missing or incompatible morphological/lexical features
- Head direction
- Clausal or non-clausal argument distinction
- Argument-adjunct distinction

Not a lot of improvements since 2015, but a lot of inconsistencies

Turkish-German code-switching treebank

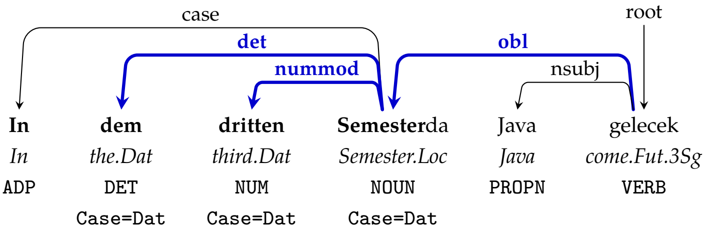
- A treebank of Turkish-German code-switching primarily developed by Özlem Çetinoğlu
- 2 184 sentences and 37 233 tokens
- 290 (surface) words were segmented
- Spoken language data, includes code-switching at sentence and word level
- Both spoken language and code-switching bring some interesting challenges
- Annotated by bilinguals, with reasonably high agreement (79% UAA, 69% LAA)
- Less agreement on Turkish tokens than German (64% vs. 72% LAA)

Code-switches



Code-switching examples

Different case assignment



case

det

nummod

obl

nsubj

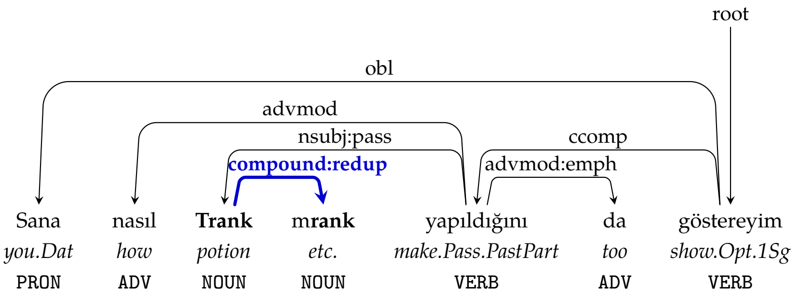
root

In	**dem**	**dritten**	**Semester**da	Java	gelecek
In	*the.Dat*	*third.Dat*	*Semester.Loc*	*Java*	*come.Fut.3Sg*
`ADP`	`DET`	`NUM`	`NOUN`	`PROPN`	`VERB`
	`Case=Dat`	`Case=Dat`	`Case=Dat`		

‘Java will start in the third semester.’

Code-switching examples

Turkish morphology applied to German



root

obl

advmod

nsubj:pass

ccomp

compound:redup

advmod:emph

Sana	nasıl	**Trank**	m**rank**	yapıldığını	da	göstereyim
you.Dat	*how*	*potion*	*etc.*	*make.Pass.PastPart*	*too*	*show.Opt.1Sg*
PRON	ADV	NOUN	NOUN	VERB	ADV	VERB

‘Let me show you also how potion et cetera is made.’

Some notes from code-switching

- Code-switching is common for many Turkic languages
- In text, it is sometimes difficult to determine which language a word belongs to. Audio recordings help
- We do not only need consistency within treebanks of the same language, but also between languages

References I

Çetinoğlu, Özlem 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)*. Paris, France: Association for Computational Linguistics, pp. 82–90. DOI: `10.18653/v1/W19-7809`. URL: `https://www.aclweb.org/anthology/W19-7809`.

Çetinoğlu, Özlem and Çağrı Çöltekin (2022). “Two languages, one treebank: building a Turkish–German code-switching treebank and its challenges”. In: *Language Resources and Evaluation*, pp. 1–35. DOI: `10.1007/s10579-021-09573-1`.

Çöltekin, Çağrı (2015). “A grammar-book treebank of Turkish”. In: *Proceedings of the 14th workshop on Treebanks and Linguistic Theories (TLT 14)*. Ed. by Markus Dickinson et al. Warsaw, Poland, pp. 35–49.

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