

# Challenges of Annotating a Code-Switching Treebank

Özlem Çetinoğlu <sup>1</sup> & Çağrı Çöltekin <sup>2</sup>

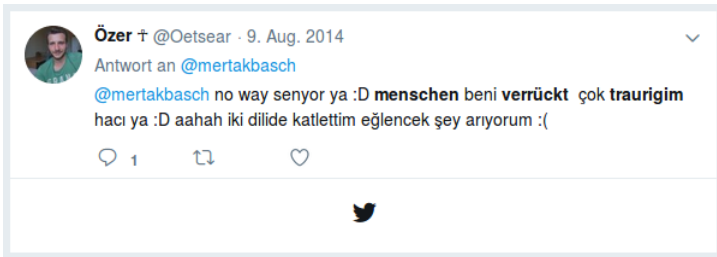
<sup>1</sup>IMS, University of Stuttgart

<sup>2</sup>SfS, University of Tübingen

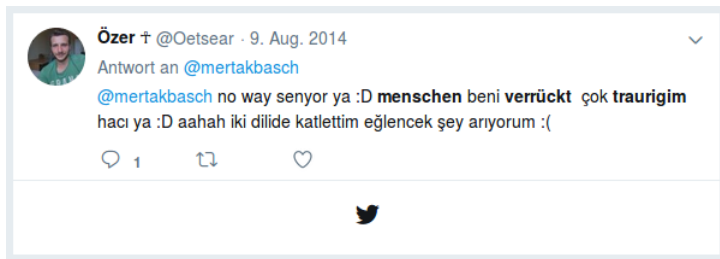
TLT – SyntaxFest

29 August 2019

# How People Communicate



# How People Communicate



- Acting multilingual, that is, mixing languages is commonly observed among multilingual speakers [Auer and Wei 2007]

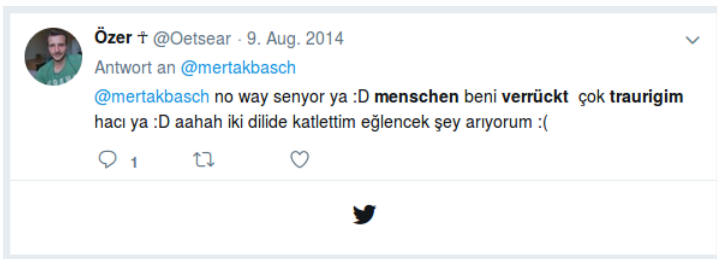
# Mixing Languages

- Extensively studied from social and linguistic perspectives  
[Poplack 1980, Myers-Scotton 1993, Muysken 2000, Auer and Wei 2007, Bullock and Toribio 2012]
- Different use of terminology
  - ▶ Code-switching (CS) for all types of mixing

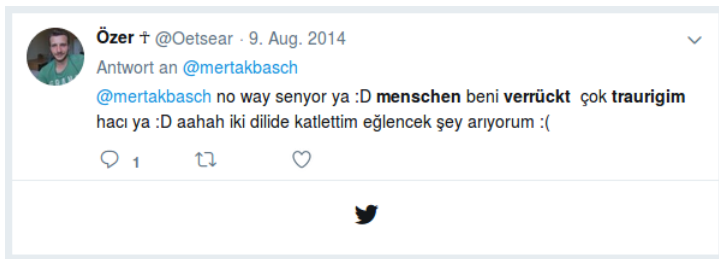
# Mixing Languages

- Extensively studied from social and linguistic perspectives  
[Poplack 1980, Myers-Scotton 1993, Muysken 2000, Auer and Wei 2007, Bullock and Toribio 2012]
- Different use of terminology
  - ▶ Code-switching (CS) for all types of mixing
- Computational studies are recent
  - ▶ Few resources, mostly shallow annotations

# Why Structural Analysis?



# Why Structural Analysis?



- Emotion Analysis

# Why Structural Analysis?

traurigim


[All](#) [Images](#) [Maps](#) [Videos](#) [News](#) [More](#) [Settings](#) [Tools](#)

About 60,700 results (0,23 seconds)


Did you mean: **traurig**

**Traurig im Kindergarten | Frage an Sozialpädagogin Christiane Schuster**  
[https://www.rund-ums-baby.de/.../Traurig-im-Kindergarten\\_96042...](https://www.rund-ums-baby.de/.../Traurig-im-Kindergarten_96042...) ▾ [Translate this page](#)  
Kindererziehung | Traurig im Kindergarten | Hallo Frau Schuster, mein Sohn (4) geht normalerweise gerne in \ seinen Kindergarten. Er ist zwar ein eher ...


**Videos**



**Manuel Neuer traurig im interview nach Niederlage gegen Südkorea ...**  
Genau  
**YouTube** - Jun 27, 2018



**Marichen sass traurig im Garten**  
Reiner Modro  
**YouTube** - Feb 6, 2012



**Toni Kroos traurig im interview nach Niederlage gegen Südkorea**  
Deutsch Aktuell  
**YouTube** - Jun 28, 2018

**Traurig im Kindergarten, Teil 2 | Frage an Sozialpädagogin Christiane ...**  
[https://www.rund-ums-baby.de/.../Traurig-im-Kindergarten-Teil-2\\_...](https://www.rund-ums-baby.de/.../Traurig-im-Kindergarten-Teil-2_...) ▾ [Translate this page](#)  
Kindererziehung | Traurig im Kindergarten, Teil 2 | Hallo Frau Schuster, ich hatte Ihnen diese Woche schon einmal geschrieben, dass mein Sohn (4) plötzlich ...

**Traurig im Kindergarten?! (Etwas länger) | NetMoms.de**  
<https://www.netmoms.de/.../Kinderbetreuung> ▾ [Translate this page](#)  
May 5, 2012 - Traurig im Kindergarten?! (Etwas länger) - Meine liebe Netmoms! Ich habe mich mal ...



# Why Structural Analysis?

traurigim

- **traurigim**  
sad.I am  
'I am sad'


[All](#) [Images](#) [Maps](#) [Videos](#) [News](#) [More](#) [Settings](#) [Tools](#)

About 60,700 results (0,23 seconds)


Did you mean: **traurig**

**Traurig im Kindergarten | Frage an Sozialpädagogin Christiane Schuster**  
[https://www.rund-ums-baby.de/.../Traurig-im-Kindergarten\\_96042...](https://www.rund-ums-baby.de/.../Traurig-im-Kindergarten_96042...) ▾ [Translate this page](#)  
Kindererziehung | Traurig im Kindergarten | Hallo Frau Schuster, mein Sohn (4) geht normalerweise gerne in \ seinen Kindergarten. Er ist zwar ein eher ...


**Videos**



**Manuel Neuer traurig im interview nach Niederlage gegen Südkorea ...**  
Genau  
YouTube - Jun 27, 2018



**Marichen sass traurig im Garten**  
Reiner Modro  
YouTube - Feb 6, 2012



**Toni Kroos traurig im interview nach Niederlage gegen Südkorea**  
Deutsch Aktuell  
YouTube - Jun 28, 2018

**Traurig im Kindergarten, Teil 2 | Frage an Sozialpädagogin Christiane ...**  
[https://www.rund-ums-baby.de/.../Traurig-im-Kindergarten-Teil-2\\_...](https://www.rund-ums-baby.de/.../Traurig-im-Kindergarten-Teil-2_...) ▾ [Translate this page](#)  
Kindererziehung | Traurig im Kindergarten, Teil 2 | Hallo Frau Schuster, ich hatte Ihnen diese Woche schon einmal geschrieben, dass mein Sohn (4) juplötzlich ...

**Traurig im Kindergarten?! (Etwas länger) | NetMoms.de**  
<https://www.netmoms.de/.../Kinderbetreuung> ▾ [Translate this page](#)  
May 5, 2012 - Traurig im Kindergarten?! (Etwas länger) - Meine liebe Netmoms! Ich brauch mal einen ...

# Why Structural Analysis?

traurigim

- **traurigim**

sad.I am  
'I am sad'

- **zenginim**

rich.I am

fakirim

poor.I am

modernim

modern.I am ...

traurigim

All

Images

Maps

Videos

News

More

Settings


Tools

About 60,700 results (0,23 seconds)


Did you mean: **traurig**

Traurig im Kindergarten | Frage an Sozialpädagogin Christiane Schuster  
[https://www.rund-ums-baby.de/.../Traurig-im-Kindergarten\\_96042...](https://www.rund-ums-baby.de/.../Traurig-im-Kindergarten_96042...) ▾ Translate this page  
Kindererziehung | Traurig im Kindergarten | Hallo Frau Schuster, mein Sohn (4) geht normalerweise gerne in \ seinen Kindergarten. Er ist zwar ein eher ...


Videos



Manuel Neuer traurig im interview nach Niederlage gegen Südkorea ...  
Genau  
YouTube - Jun 27, 2018



Marichen sass traurig im Garten  
Reiner Modro  
YouTube - Feb 6, 2012



Toni Kroos traurig im interview nach Niederlage gegen Südkorea  
Deutsch Aktuell  
YouTube - Jun 28, 2018

Traurig im Kindergarten, Teil 2 | Frage an Sozialpädagogin Christiane ...  
[https://www.rund-ums-baby.de/.../Traurig-im-Kindergarten-Teil-2\\_...](https://www.rund-ums-baby.de/.../Traurig-im-Kindergarten-Teil-2_...) ▾ Translate this page  
Kindererziehung | Traurig im Kindergarten, Teil 2 | Hallo Frau Schuster, ich hatte Ihnen diese Woche schon einmal geschrieben, dass mein Sohn (4) jurplötzlich ...

Traurig im Kindergarten?! (Etwas länger) | NetMoms.de  
<https://www.netmoms.de/.../Kinderbetreuung> ▾ Translate this page  
May 5, 2012 - Traurig im Kindergarten?! (Etwas länger) - Meine liebe Netmami! Ich brauch mal einen ...

# Why Structural Analysis?

- (1) no way senyor :D **menschen** beni **verrückt** çok **traurig***im*  
mister people me crazy very sad. I am  
'no way mister, **people** make me **crazy**, *I am* very **sad**'

# Annotating a Code-Switching Treebank

- Speech data collection
- Transcription and normalisation
- POS and morphology annotation
- Dependency annotation

# Annotating a Code-Switching Treebank

- Speech data collection
  - Transcription and normalisation
  - POS and morphology annotation
  - Dependency annotation
- 
- Follows Universal Dependencies
    - ▶ so far: UD v2.3 + Turkish IMST, German GSD
    - ▶ next: UD v2.4 + also other Turkish and German treebanks
- 
- Complete treebank: 2000 sentences
    - ▶ Observations: on 1/3 of the corpus

# Challenges

- Annotation Differences in Individual Languages
  - ▶ Titles
  - ▶ Copula
  - ▶ Case

# Challenges

- CS-specific Issues
  - ▶ Double case marking
  - ▶ Bilingual light verb constructions
  - ▶ Translation pairs
  - ▶ Bilingual *m*-reduplication

# Challenges

- Issues Related to Spoken Language
  - ▶ Appositions
  - ▶ Dislocation
  - ▶ Clausal discourse elements

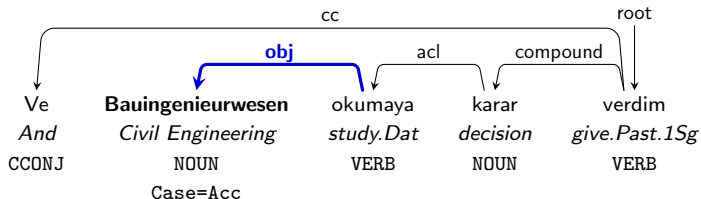


# Challenges

- Annotation Differences in Individual Languages
  - ▶ Titles
  - ▶ Copula
  - ▶ Case
- CS-specific Issues
  - ▶ Double case marking
  - ▶ Bilingual light verb constructions
  - ▶ Translation pairs
  - ▶ Bilingual *m*-reduplication
- Issues Related to Spoken Language
  - ▶ Appositions
  - ▶ Dislocation
  - ▶ Clausal discourse elements

# Annotation Differences in Individual Languages

- Case

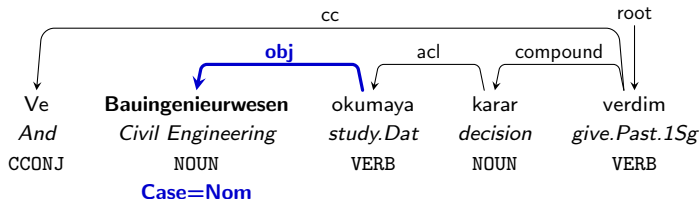


'And I decided to study Civil Engineering'

- German annotation standards: Object is accusative

# Annotation Differences in Individual Languages

- Case

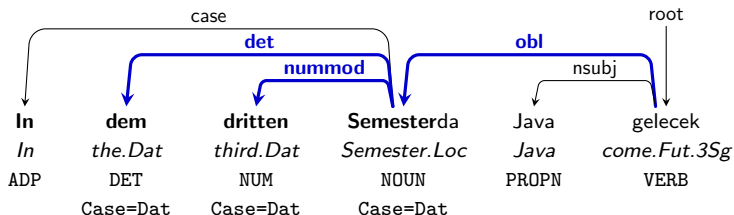


'And I decided to study Civil Engineering'

- German annotation standards: Object is accusative
- Turkish annotation standards: Object is accusative only when overt

# CS-specific Issues

- Double case marking

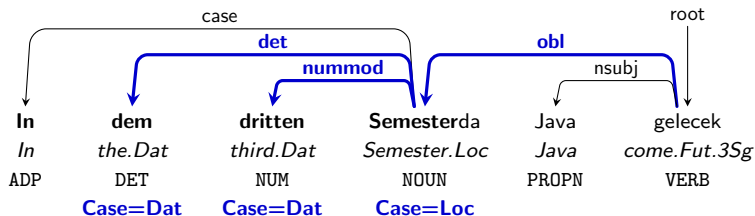


'Java will start in the third semester.'

- Case agreement within the German NP

# CS-specific Issues

- Double case marking

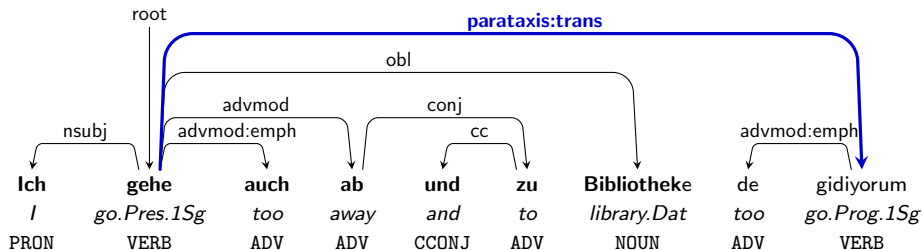


‘Java will start in the third semester.’

- Case agreement within the German NP
- Additional Turkish case marker

# CS-specific Issues

- Translation pairs

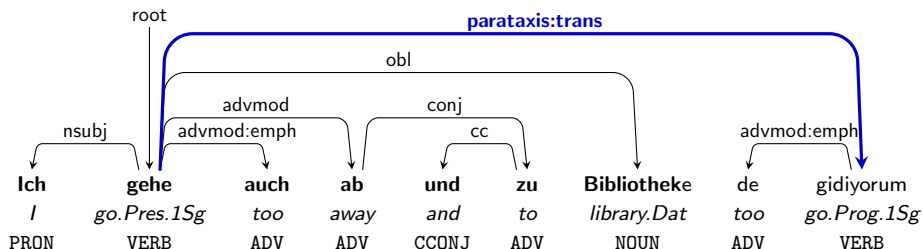


'I also go to the library now and then.'

- Uttering a word, phrase or clause in one language and repeating it as a translation in the other language

# CS-specific Issues

- Translation pairs



'I also go to the library now and then.'

- Introduce the subtype `parataxis:trans`

# CS-specific Issues

- Bilingual *m*-reduplication

(2) Çay **m**ay içer misin?  
*Tea etc. drink.Aor Ques.2Sg*

‘Would you like to drink tea and the like?’

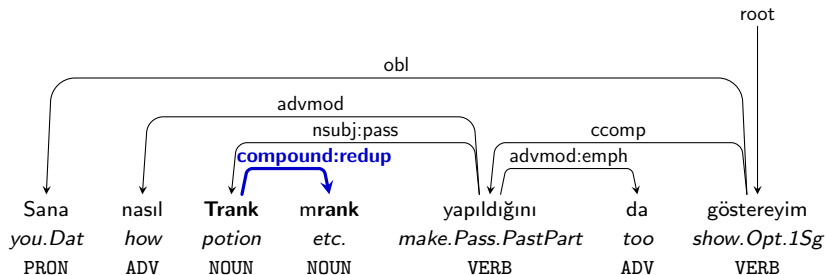
- Turkish *m*-reduplication

- ▶ a word starts with a consonant → reduplicate and replace the consonant with an *m*
- ▶ a word starts with a vowel → reduplicate and add an *m* prefix



# CS-specific Issues

- Bilingual *m*-reduplication

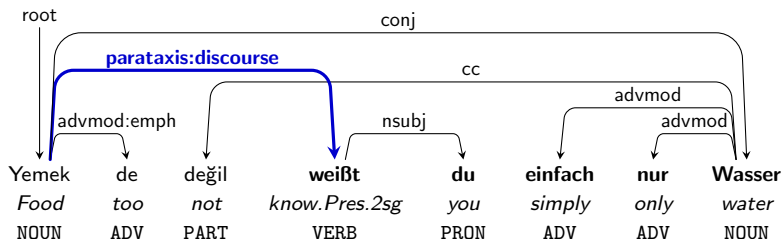


'Let me show you also how potion et cetera is made.'

- Follow the Turkish dependency subtype

# Issues Related to Spoken Language

- Clausal discourse elements



'It is not food, you know, just water.'

- Follow UD spoken treebanks: discourse information via the subtype [Dobrovoljc and Nivre 2016, Gerdes and Kahane 2017, Courtin et al. 2018]

# Summary

- Building a Turkish-German UD treebank
  - ▶ Work in progress

# Summary

- Building a Turkish-German UD treebank
  - ▶ Work in progress
- Observing challenges
  - ▶ Differences between and within languages/treebanks/guidelines
  - ▶ New syntactic structures due to CS
  - ▶ Frequent spoken language phenomena

# Summary

- Building a Turkish-German UD treebank
  - ▶ Work in progress
- Observing challenges
  - ▶ Differences between and within languages/treebanks/guidelines
  - ▶ New syntactic structures due to CS
  - ▶ Frequent spoken language phenomena
- Opening the annotations up to discussion
  - ▶ Inter and intra-language treebank comparisons
  - ▶ Annotations from other treebanks
  - ▶ Suggestions for new dependency subtypes

Thanks!

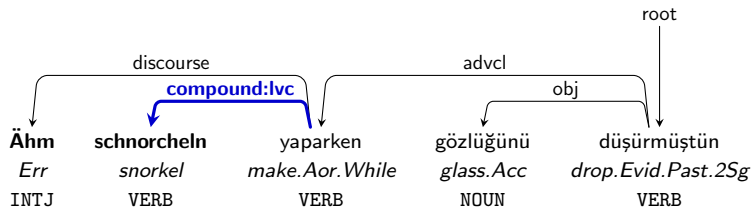
Questions?

# References I

- Auer, P. and Wei, L. (2007). *Handbook of multilingualism and multilingual communication*, volume 5. Walter de Gruyter.
- Bullock, B. E. and Toribio, A. J. (2012). *The Cambridge handbook of linguistic code-switching*. Cambridge University Press.
- Courtin, M., Caron, B., Gerdes, K., and Kahane, S. (2018). Establishing a language by annotating a corpus: the case of Naija, a post-creole spoken in Nigeria. In *Proceedings of the Workshop on Annotation in Digital Humanities*, pages 7–11.
- Dobrovoljc, K. and Nivre, J. (2016). The Universal Dependencies treebank of spoken Slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.
- Gerdes, K. and Kahane, S. (2017). Trois schémas d'annotation syntaxique en dépendance pour un même corpus de français oral : le cas de la macrosyntaxe. In *Atelier sur les corpus annotés du français (ACor4French)*.
- Muysken, P. (2000). *Bilingual Speech: A Typology of Code-Mixing*.
- Myers-Scotton, C. (1993). *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.
- Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en Español: toward a typology of code-switching. *Linguistics*, 18(7-8):581–618.

# CS-specific Issues

- Bilingual light verb constructions



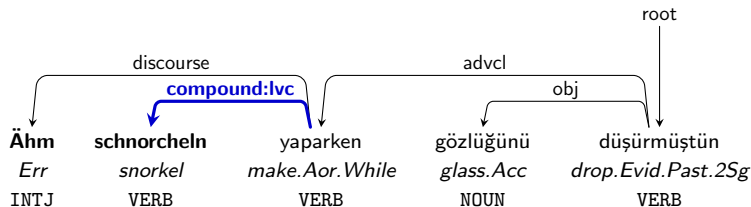
'While you were snorkelling you had dropped your glass.'

- Also in other language pairs: Turkish–Dutch, Turkish–English, Hindi–English



# CS-specific Issues

- Bilingual light verb constructions



'While you were snorkelling you had dropped your glass.'

- Similar to noun-light verb constructions common in Turkish  
e.g. *yardım etmek* lit. 'help do' – 'to help'