UD for German Poetry

Stefanie Dipper & Ronja Laarmann-Quante

Department of Linguistics
Ruhr University Bochum

TLT 2024
University of Hamburg
Dec 5–6, 2024

RUB

Overview of the talk

- Background
- UD / UD for German
- UD for German poetry
- Evaluation of PoeTree.de

Some background

Why are we interested in Universal Dependencies (UD)?

- Work on **syntactic change** in German
  - Middle High German to Modern German (1050 – today)
- Work on **language and literacy acquisition**
  - books/poems written for children that use specific linguistic patterns
  - e.g. rhymes, repetitions of words, of syntactic structures, . . .
- This work: pilot study on German poetry for adults

RUB

Overview

Universal Dependencies

(de Marneffe, Manning, Nivre, and Zeman 2021)

- Universal scheme for syntactic dependencies for any language
- This means: compromises for many languages
- Focus on semantics: function words are dependents from content words
- Starting point: Stanford dependencies for English

Benefit from using Universal Dependencies

- Annotations are easier …
  - to understand in foreign language treebanks
  - to compare between different languages
  - to produce (support by tools)
  - to produce with good performance (more training data)

The official German UD scheme

Problems with the German scheme

1. Terminology
   - dative, genitive objects: analyzed as `obl:arg`
   - secondary accusative objects: `iobj`
   - negated article *kein* 'no': `advmod` (as of 02/2023)
2. Missing key distinctions
   - no distinction between infinitival subjects and objects: both `xcomp`
   - no distinction between different types of expletives
   - prepositional phrases (arguments and modifiers): all analyzed as `obl`

Map other schemes to Universal Dependencies

- Reviewer: don't use UD if you want to encode all these language-specific properties; use traditional schemes like TIGER (Albert et al. 2005) or TüBa (Telljohann et al. 2012) instead
- But: we want to profit from the UD resources
- But: our data would be "lost" for UD
  - annotations could be mapped to UD scheme
  - but any mapping carries the risk of errors
  - ideally, mappings retain all information but often not

RUB

Universal Dependencies: extensions for German

(Dipper, Haiber, Schröter, Wiemann, and Brinkschulte 2024a)

Our solutions:

1. Terminology
   - dative, genitive objects: analyzed as `iobj`, following Zeman (2017)
2. Missing key distinctions $\rightarrow$ we define new subtypes in the form `universal:extension`, e.g. `xcomp:subj`

Applications of the extended scheme

- German poetry $\rightarrow$ see below
- Middle High German (MHG)
  - treebank of almost 29K tokens
  - IAA: $\alpha = .85$ (Skjærholt 2014)
  - below: some constructions specific to MHG

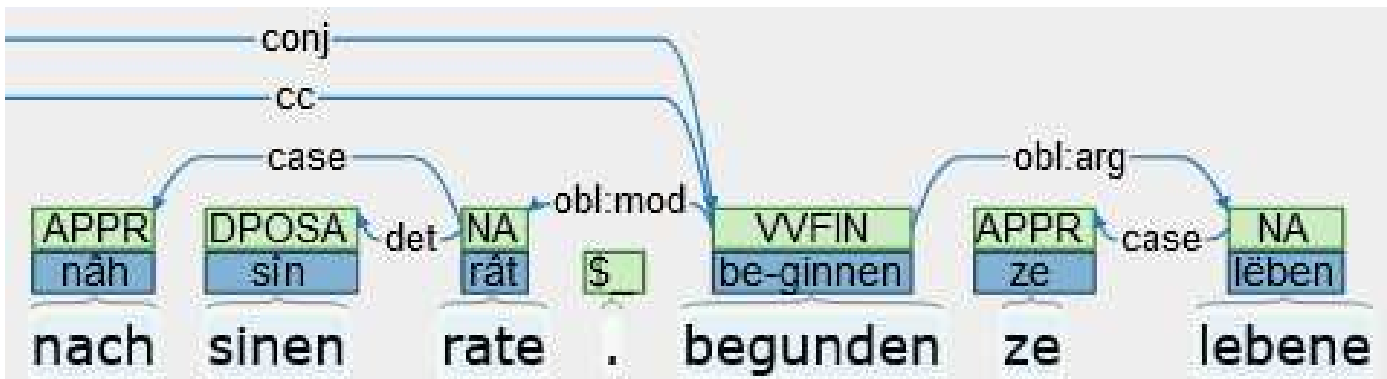
MHG: *to*-infinitive

First: preposition + gerund (`case` + noun)

Later: infinitive marker + infinitive (`mark` + verb)

Example

*nach sinen rate begunden **ze lebene***
according to his advice began **to living.DAT**
'began **to live** by his advice'



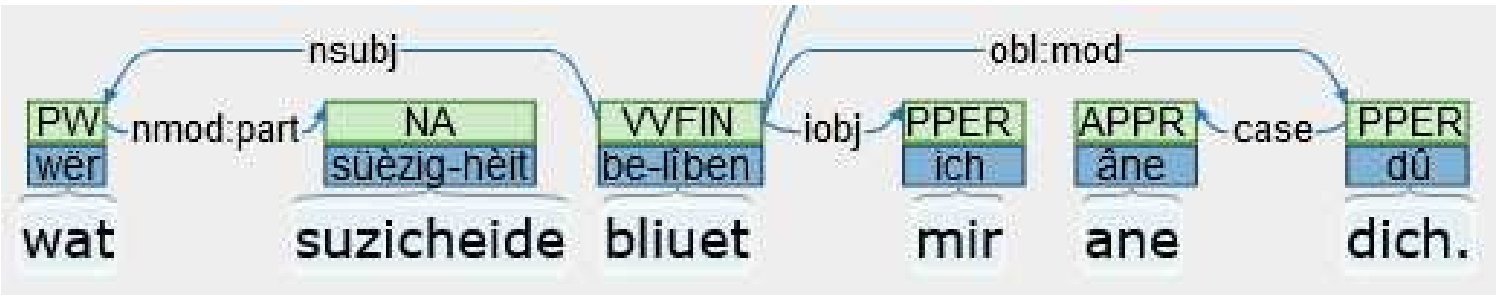
MHG: partitive

First: quantity + substance noun (gen) (quant + `nmod:part`)

Later: determiner + head noun (`det` + noun)

Example

wat suzicheide *bliuet mir ane dich.*
what sweetness.GEN.SG remains me without you
‘**what sweetness** remains to me without you’



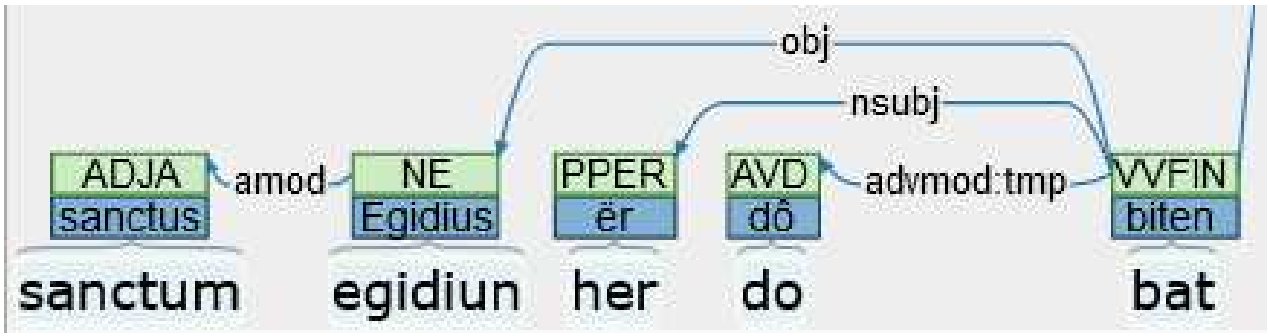
MHG: proper nouns

First: ordinary adjective + name

Later: complex proper name

Example

sanctum egidiun *her do bat*
holy/saint Aegidius he then asked
‘he then asked **Saint Aegidius**’

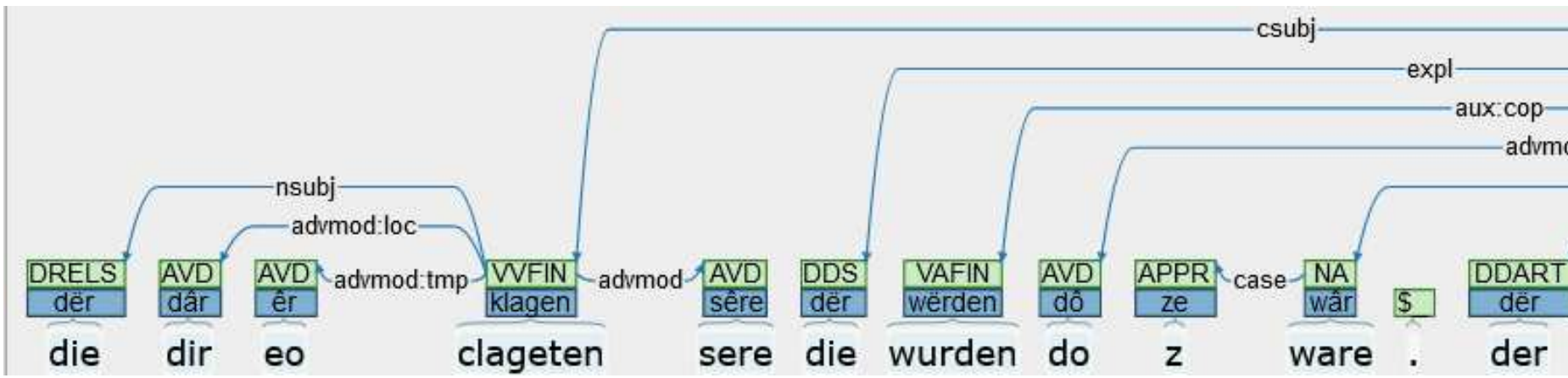


MHG: dislocation

- Significantly more instances of dislocation in MHG
- Presumably because of more oral style
- UD: function of the dislocated element is not recorded

Example

*[**die dir eo clageten sere**] **die** wurden do z ware der gotelichen werke uro*
[**who there before complained a lot**] **they** became then indeed the divine works glad
'[**those who complained there before**], **they** became glad indeed of the works of God'



Overview

NLP for poetry

- **Repetitive structures and patterns**: support memorization
  - e.g. meter and the rhyme scheme
  - most computational approaches focus on these features
- **Syntactic repetitions**
  - e.g. ‘parallel couplets’: semantic or syntactic correspondences between two lines
  - Lee and Kong (2012): poems in classical Chinese
- **Unusual word order**
  - e.g. enjambment: elements of a syntactic phrase are spread over two lines
  - Ruiz Fabo et al. (2017) for diachronic Spanish
  - Hussein et al. (2018) for German (read-out) poetry

RUB

UD for German poetry

- A first pilot study
- Poetry-specific challenges
- A (small) quantitative evaluation

PoeTree corpus

- Our data: taken from the PoeTree corpus
- **PoeTree corpus** (Plecháč et al. 2024)
  - more than 330K poems with 89M tokens from 10 European languages
  - all poems have been annotated automatically with UD-style dependencies using UDPipe 2.0 (Straka et al. 2016)
  - only the annotations of the Czech-language subcorpus have already been evaluated (Cinková et al. 2024) (more details below)

Our data

- Random selection of 20 poems from PoeTree.de, the German subcorpus of PoeTree, mainly from 19th century
- Manual annotation of dependency relations with INCEpTION (Klie et al. 2018)
- Each poem was annotated once, by one of the authors, and difficult cases were discussed together

	Mean ± SD	**Total**
Tokens	108.1 ± 92.8	2,162
Lines	15.7 ± 15.1	314
Stanzas	4.9 ± 5.7	97

Poetry – reviews – news

- Comparison of label distribution with two subsets of the GSD treebank: modern news and reviews
- Pairwise Spearman's rank correlation coefficient *r*

	News	**Reviews**	**Poetry**
News	1		
Reviews	0.94	1	
Poetry	0.76	0.83	1

→ Poetry more similar to reviews than to news; e.g. news have fewer coordinations

Overview

Problematic spellings and tokenization

- Capitalization of first word in line: often leads to incorrect part-of-speech tags (as noun)
- Elisions with apostrophe: often incorrectly tokenized as three tokens (e.g. *heil'gen* = *heiligen* 'holy')
- UD-specific: contracted preposition + article (*aufs* 'at the') are split (into *auf das*), which often changes the meaning

Example

Und was dir fehlschlug, hoffe stets ***aufs neu'***
'And what you have failed, always hope **anew**'
(and not: 'always hope **for the new**')

Incorrect sentence boundaries

- Non-standard punctuation + unusual capitalization: often leads to incorrect sentence boundaries

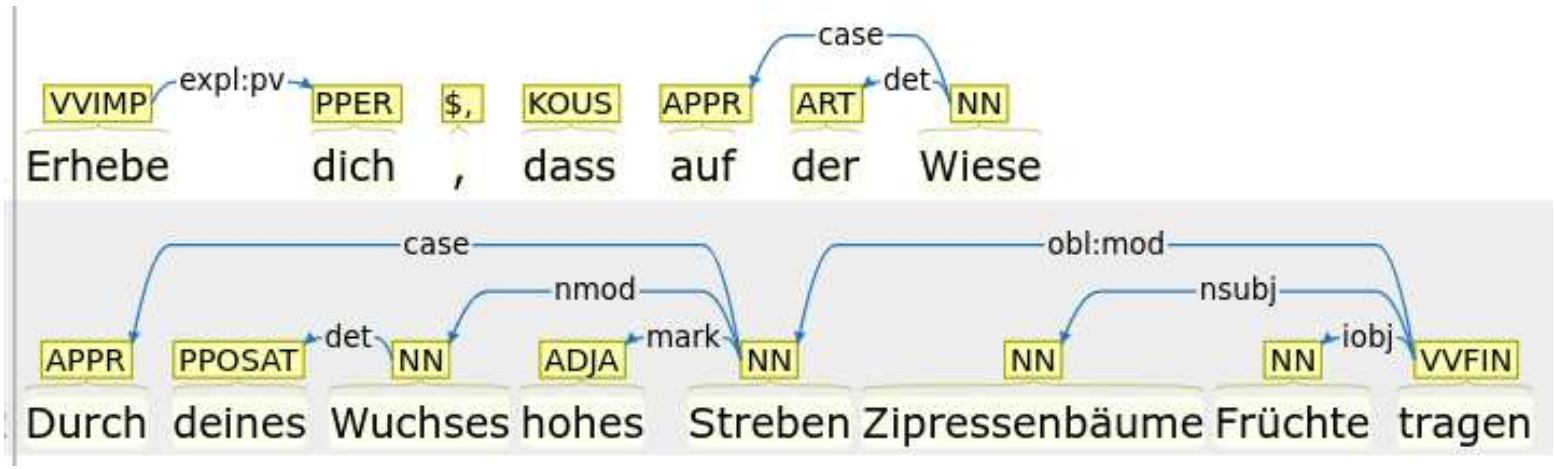
Example

Geduld!
Geduld! *– die ew'gen Sterne gehn Doch ihren Pfad.*
'**Patience! Patience!** – the eternal stars go but their path.'

Incorrect sentence boundaries

Example

Erhebe dich, dass auf der Wiese
Durch deines Wuchses hohes Streben Zipressenbäume Früchte ***tragen***
'Arise, so that in the meadow through your growth's high aspiration cypress trees **bear** fruit'



Incorrect sentence boundaries

- Incorrect sentence boundaries tear apart dependency-related phrases
- In difficult passages, incorrect sentence boundaries can distort the meaning
- Dependency analysis should be carried out first, on the basis of which the sentence boundaries would then be determined
- Alternatively, parsing and sentence boundary detection could be performed in parallel

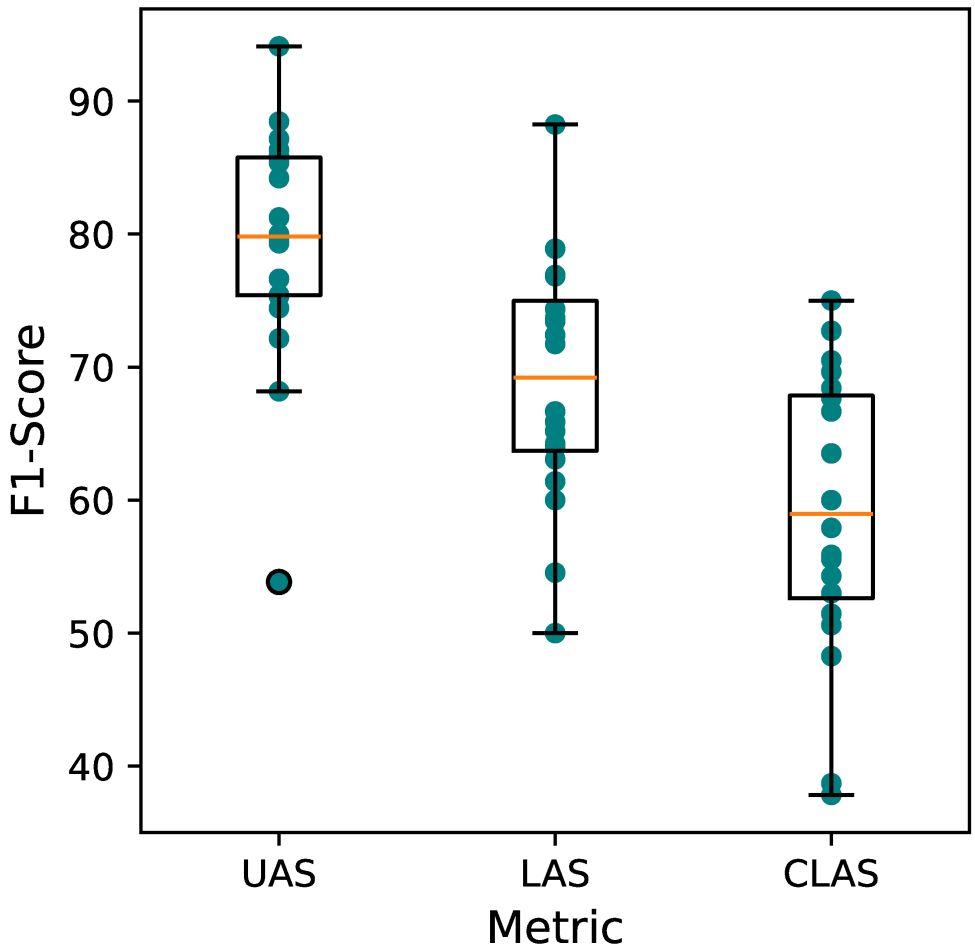
Quantitative evaluation: poetry vs. others / German vs. Czech

	PT.de	GSD	PT.cz	PDT
Language	**de**	**de**	cz	cz
Text type	**poetry**	news/reviews/web	**poetry**	news
UAS	**79.6**	82.8	85.0	95.0
LAS	**68.9**	78.2	79.7	93.6
CLAS	**59.2**	–	–	–

- CLAS: Content-Word Labeled Attachment Score (Zeman et al. 2017)
- Parses from UDPipe 2.0
  - PT.de: our evaluation
  - GSD/PDT: evaluation by Straka (2018)
  - PT.cz: evaluation by Plecháč et al. (2024)

UAS / LAS / CLAS

Distribution of UAS, LAS and CLAS scores



Error analysis

- Top five confused dependency labels between manual and automatic annotations

Manual	PoeTree	F1	count
advcl	ccomp	0.39	7
parataxis	conj	0.28	26
obl	nmod	0.28	25
expl	obj	0.15	13
iobj	obj	0.13	11

- `advcl` vs. `ccomp` also problematic in Middle High German (Dipper et al. 2024b)
- `iobj` vs. `obj`: differing guidelines (all dative objects vs. ditransitive only)

Conclusion

- Automatic preprocessing problematic
- Automatic parses not yet reliable enough
- Incorrect sentence boundaries are a major problem
- Need more manual annotations

Paper and Corpus

- **Paper**:
  Stefanie Dipper Dipper & Ronja Laarmann-Quante (2024). UD for German poetry. In *Proceedings of NLP4DH*, Miami, USA, pp. 177–188.
- **Corpus** available at:
  `https://gitlab.ruhr-uni-bochum.de/vamos-cl/ud-for-german-poetry`

Thank you!

RUB

References I

Albert, S. et al. (2005).
TIGER-Annotationsschema.
Technical report, Universität Potsdam, Universität Saarbrücken, Universität Stuttgart.

Cinková, S., P. Plecháč, and M. Popel (2024, June).
Rhymes and syntax: A morpho-syntactic analysis of Czech poetry.
Primerjalna književnost 47(2).

de Marneffe, M.-C., C. D. Manning, J. Nivre, and D. Zeman (2021, 07).
Universal Dependencies.
Computational Linguistics 47(2), 255–308.

Dipper, S., C. Haiber, A. M. Schröter, A. Wiemann, and M. Brinkschulte (2024a).
Universal Dependencies: Extensions for modern and historical German.
In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia, pp. 17101–17111.

RUB

References II

Dipper, S., C. Haiber, A. M. Schröter, A. Wiemann, and M. Brinkschulte (2024b).
Universal Dependencies: Extensions for modern and historical German.
In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia, pp. 17101–17111.

Hussein, H., B. Meyer-Sickendiek, and T. Baumann (2018).
Automatic detection of enjambment in German readout poetry.
In *Proceedings of Speech Prosody, 2018, Poznán*, pp. 329–333.

Klie, J.-C., M. Bugert, B. Boullosa, R. E. de Castilho, and I. Gurevych (2018).
The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation.
In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pp. 5–9. Association for Computational Linguistics.

RUB

References III

Lee, J. and Y. H. Kong (2012, June).
A dependency treebank of classical Chinese poems.
In E. Fosler-Lussier, E. Riloff, and S. Bangalore (Eds.), *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montréal, Canada, pp. 191–199. Association for Computational Linguistics.

Plecháč, P., S. Cinková, R. Kolár, A. Šeļa, M. De Sisto, L. Nugues, T. Haider, and N. Kočnik (2024, September).
PoeTree: Poetry treebanks in Czech, English, French, German, Hungarian, Italian, Portuguese, Russian, Slovenian and Spanish.
Research Data Journal for the Humanities and Social Sciences, 1–17.

Ruiz Fabo, P., C. Martínez Cantón, T. Poibeau, and E. González-Blanco (2017, August).
Enjambment detection in a large diachronic corpus of Spanish sonnets.
In B. Alex, S. Degaetano-Ortlieb, A. Feldman, A. Kazantseva, N. Reiter, and S. Szpakowicz (Eds.), *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, Vancouver, Canada, pp. 27–32. Association for Computational Linguistics.

RUB

References IV

Skjærholt, A. (2014, June).
A chance-corrected measure of inter-annotator agreement for syntax.
In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, Maryland, pp. 934–944. Association for Computational Linguistics.

Straka, M. (2018, October).
UDPipe 2.0 prototype at CoNLL 2018 UD shared task.
In D. Zeman and J. Hajič (Eds.), *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Brussels, Belgium, pp. 197–207. Association for Computational Linguistics.

Straka, M., J. Hajič, and J. Straková (2016, May).
UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing.
In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, pp. 4290–4297. European Language Resources Association (ELRA).

RUB

References V

Telljohann, H., E. W. Hinrichs, S. Kübler, H. Zinsmeister, and K. Beck (2012). *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Germany: Seminar für Sprachwissenschaft, Universität Tübingen.

Zeman, D. (2017). Core arguments in Universal Dependencies. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, Pisa, Italy, pp. 287–296.

Zeman, D., M. Popel, M. Straka, J. Hajič, J. Nivre, F. Ginter, J. Luotolahti, S. Pyysalo, S. Petrov, M. Potthast, F. Tyers, E. Badmaeva, M. Gokirmak, A. Nedoluzhko, S. Cinková, J. Hajič jr., J. Hlaváčová, V. Kettnerová, Z. Urešová, J. Kanerva, S. Ojala, A. Missilä, C. D. Manning, S. Schuster, S. Reddy, D. Taji, N. Habash, H. Leung, M.-C. de Marneffe, M. Sanguinetti, M. Simi, H. Kanayama, V. de Paiva, K. Droganova, H. Martínez Alonso, Ç. Çöltekin, U. Sulubacak, H. Uszkoreit, V. Macketanz, A. Burchardt, K. Harris, K. Marheinecke, G. Rehm, T. Kayadelen, M. Attia, A. Elkahky, Z. Yu, E. Pitler, S. Lertpradit, M. Mandl, J. Kirchner, H. F. Alcalde, J. Strnadová, E. Banerjee, R. Manurung, A. Stella, A. Shimada, S. Kwak, G. Mendonça, T. Lando, R. Nitisaroj, and J. Li (2017, August).

RUB

References VI

CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies.
In J. Hajič and D. Zeman (Eds.), *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Canada, pp. 1–19. Association for Computational Linguistics.

RUB