Dependencies over Times and Tools (DoTT)

Andy Lücking, Giuseppe Abrami, Leon Hammerla, Marc Rahn, Daniel Baumartz, Steffen Eger, Alexander Mehler

Goethe University Frankfurt, Text Technology Lab – Senckenberg Society for Nature Research, Leibniz Institution for Biodiversity and Earth System Research – Bielefeld University

LREC-COLING 2024

Question

Based on the examples of English and German, we investigate to what extent parsers trained on modern variants of these languages can be transferred to older language levels without loss. The DoTT corpus will be available at `https://github.com/texttechnologylab/DoTT`.

Corpora

	GerParCor (Abrami et al., 2022)
	includes Bundestag and DEUParl (Walter et al., 2021)
link	`https://github.com/texttechnologylab/GerParCor`
description	Plenary protocols of all German-speaking parliaments (Austrian National Council, Swiss National Council, Liechtenstein National Parliament, German Bundestag, German Bundesrat, and the 16 German national parliaments).
time / lang.	1867–2021 / DE
	coha
link	`https://www.english-corpora.org/coha/`
description	Corpus of Historical American English, 400 million words / 107,000 texts
time / lang.	1810–2009 (acc. to website) / EN
	hansard
link	`https://www.english-corpora.org/hansard/`
description	nearly every speech given in the British Parliament (about 1.6 billion words total acc. to website)
time / lang.	1803–2005 / EN
	dta
link	`https://www.deutschestextarchiv.de`, `http://media.dwds.de/dta/download/dta_kernkorpus_2020-07-20.zip`
description	Kernel corpus of the German Text Archive (*Deutsches Textarchiv*), version from July 20, 2020, 1,472 texts, 359M; Belle lettres (552 texts, 92M), functional literature (266 texts, 71M), science (654 texts, 198M)
time / lang.	1600–1899 (1600–1699: 237 texts, 60M; 1700–1799: 526 texts, 122M; 1800–1899: 689 texts, 167M) / DE
	BIOfid
link	`https://www.biofid.de/de/\#digital-collection` (CC BY-NC-SA 4.0)
description	German botanical and biodiversity texts from the collections of the library of Goethe University Frankfurt, accessed via the specialized information service BIOfid (Driller et al., 2020).
time / lang.	since 1753 / DE

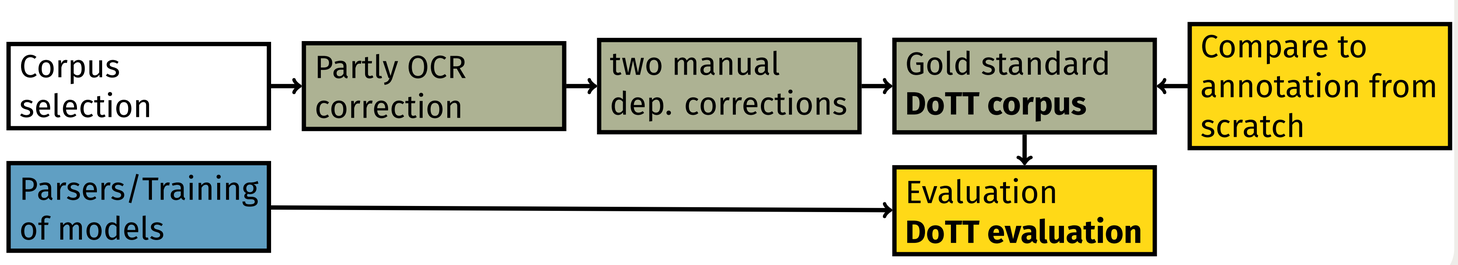
Goldstandard Corpus

lang	corpus	time	dep tag	#sent	#bckt
DE	parl	1895–1942	TIGER	168	96
DE	parl	1895–1942	UD	161	93
DE	bio	1753–today	TIGER	270	—
EN	hans	1803–2005	UD	107	—
			sums:	706	189

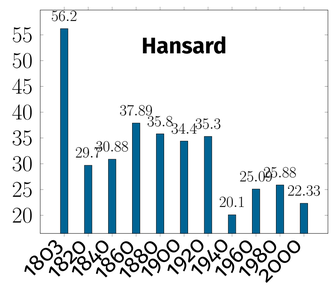
Models on older/newer sentences

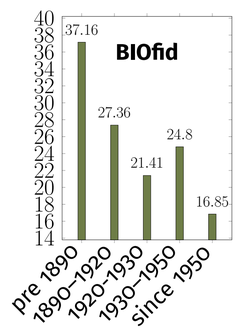
Model	UAS old / new	LAS old / new
biaffine_dep_de_tiger	88.64 / 91.03	85.79 / 88.12
crf_dep_de_tiger	87.17 / 89.36	83.91 / 86.09
crf2o_dep_de_tiger	86.64 / 89.35	83.75 / 86.13
biaffine_dep_en_gum	88.64 / 91.26	85.32 / 87.01
crf_dep_en_gum	88.54 / 89.64	85.46 / 85.29
crf2o_dep_en_gum	88.21 / 89.71	84.82 / 85.43
biaffine_dep_de_gsd	89.18 / 89.68	79.47 / 83.13
crf_dep_de_gsd	88.46 / 89.81	79.35 / 83.66
crf2o_dep_de_gsd	88.66 / 90.49	80.68 / 83.93
Stanford UD	89.53 / 90.67	86.03 / 86.86
average	88.78 / 90.36	84.60 / 86.14

Workflow



Results I





Old texts have slightly longer sentences, which affects parsers (see Models on older/newer sentences), but not statistically significant

Results II

model	LAS doc	LAS macro	UAS doc	UAS macro
biofid				
biaffine tiger	85.36	83.64	88.61	87.02
crf tiger	83.52	81.56	86.95	85.35
cfr2o tiger	83.28	80.98	86.73	84.84
parliamentary ud				
biaffine gsd	82.89	83.67	90.51	91.42
crf gsd	83.54	84.19	90.61	91.49
crf2o gsd	83.78	84.24	90.83	91.39
Stanford UD	85.33	86.67	90.93	92.08
parliamentary tiger				
biaffine tiger	88.54	88.38	91.16	91.13
crf tiger	87.57	87.66	90.46	90.78
crf2o tiger	87.13	87.15	89.94	90.14
hansard				
biaffine gum	85.93	86.45	89.59	89.25
crf gum	85.40	85.54	88.94	88.24
crf2o gum	85.04	85.47	88.75	88.98
Stanford UD	86.76	88.06	87.70	90.91

With an average LAS (sentence) of 85.29, parsers show to work quite reliable, even on older texts.

Results III

- IAA between the two raters was about 0.9 (see appendix of paper).
- However, human annotators may be biased by trusting parser annotations (Fort and Sagot, 2010).
- In order to assess the magnitude of this bias, 46 sentences from the parliamentary have been drawn and manually annotated by hand.
- LAS numbers of about 0.74 are slightly worse than those of dependency tools, confirming the bias to believe the machine.

Discussion

- Dependency-length studies on older texts are feasible using modern dependency parsers.
- But do they recognize syntactic change?
- see examples 👉

Examples

(1) a. Quene Ester looked never with swich an eye.

(Chaucer, *Merchant's Tale*, line 1744, from the end of 14th century, cited in Kroch 1989, p. 226)

b. Quene Ester never looked with swich an eye.

(2) *dhazs ir chihoric uuari gote*
dass er gehorsam war Gott
that he obedient was God

"dass er gehorsam war gegenüber Gott" / *that he was obedient to God*

All examples are parsed by dependency parser, but (1-a) and (2) are ungrammatical nowadays due to language change.

References

Abrami, Giuseppe et al. (2022). "German Parliamentary Corpus (GerParCor)". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 1900–1906. URL: `https://aclanthology.org/2022.lrec-1.202`.

Driller, Christine et al. (2020). "Fast and Easy Access to Central European Biodiversity Data with BIOfid". In: *Biodiversity Information Science and Standards*. Vol. 4. BISS, e59157. DOI: `10.3897/biss.4.59157`.

Fort, Karen and Benoît Sagot (July 2010). "Influence of Pre-annotation on POS-tagged Corpus Development". In: *The Fourth ACL Linguistic Annotation Workshop*. Uppsala, Sweden, pp. 56–63. URL: `https://hal.archives-ouvertes.fr/hal-00484294`.

Kroch, Anthony S. (1989). "Reflexes of grammar in patterns of language change". In: *Language Variation and Change* 1.3, pp. 199–244. DOI: `10.1017/S0954394500000168`.

Walter, Tobias et al. (2021). "Diachronic Analysis of German Parliamentary Proceedings: Ideological Shifts through the Lens of Political Biases". In: *ACM/IEEE Joint Conference on Digital Libraries*. JCDL'21, pp. 51–60.