Tracing Syntactic Change in the Scientific Genre: Two Universal Dependency-parsed Diachronic Corpora of Scientific English and German

Marie-Pauline Krielke*, Luigi Talamo*, Mahmoud Fawzi*, Jörg Knappen*
*Saarland University

GOALS

1. Improve parsing quality by pre-processing historical input data (PIPELINE)

DTAW

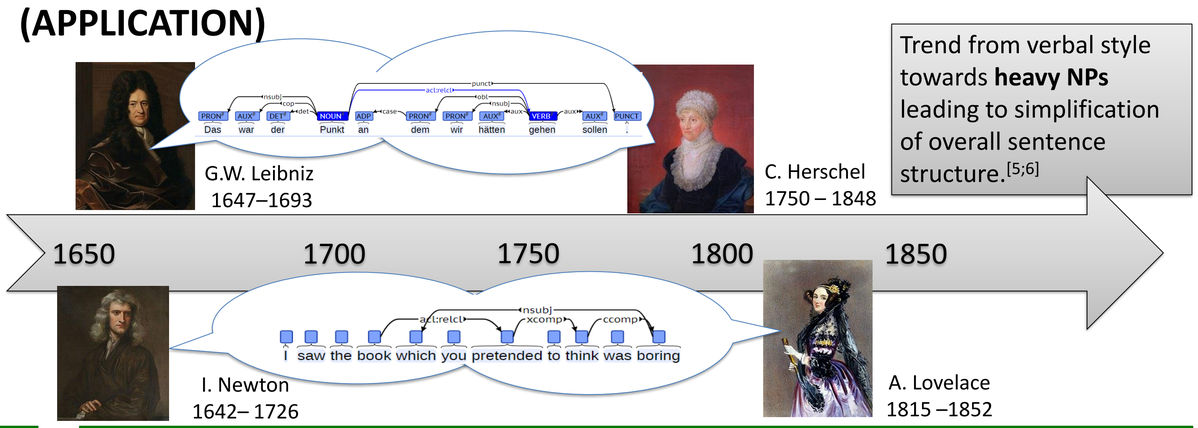
Scientific texts from the *Deutsches Textarchiv* [2] (1650 – 1899) with ca. 82 million tokens.
Wordforms canonicalized with CAB, tokenized with DTA-Tokwrap[3].
https://www.deutschestextarchiv.de

Royal Society Corpus

RSC [1] : texts from the Philosophical *Transactions* and *Proceedings* of the Royal Society of London (1665 – 1899) ca. 32 million tokens with standard linguistic annotation (+ normalization of historical word forms, VARD [4]).
http://fedora.clarin-d.uni-saarland.de/rsc_v6/

Links to parsed corpora:
RSC_UD-Parsed_1.0: http:hdl.handle.net/21.11119/0000-000A-A556-B
DTAW_UD-Parsed_1.0: http:hdl.handle.net/21.11119/0000-000A-A555-C

2. Trace syntactic change in EN and DE scientific discourse (APPLICATION)



PIPELINE

1. Normalization of Historical Data

Replacing the formerly common virgule (slash) by the analogous comma:
„Wann jemand etwas seinem Nächsten zum Besten aufrichtig heraus gibt / so gering es auch ist / billig zu Dank soll angenommen werden.“ → Wann jemand etwas seinem Nächsten zum Besten aufrichtig heraus gibt, so gering es auch ist, billig zu Dank soll angenommen werden.

2. Extraction of “good sentences“ (GS)

Detection of non-sentential constructions (“bad sentences“ – BS)

- Sentences beginning in lower case and the preceding sentence (incomplete)
- Sentences with less than 8 tokens (too short)
- Sentences lacking a verb (verbless)
- Foreign-language sentences (foreign)

3. UD-parsing

Parser: UDPipe 1.0, UD Models (2.5): GUM (EN), GSD (DE)

- Input tokens: normalized word forms
- Preservation of sentence splitting and tokenization

4. Evaluation

Sample 100 GS and BS (20 / 50 years period, e.g., 1650–1699)

4.1 Parsability of a sentence: grammatically interpretable structure

- Accepted structures: title-like noun phrases and dates
- Excluded structures: sentences in other language, fragments without grammatical structure, i.e., equations, abbreviations

4.2 Roots: number and accuracy of sentence roots

4.3 Parsing accuracy: UD label and syntactic head

- Overall parsing accuracy ~ 80% for both corpora
- Significant improvement (Ø 23%) of accuracy for GS vs. BS
- Stable accuracy of GS over time

APPLICATION

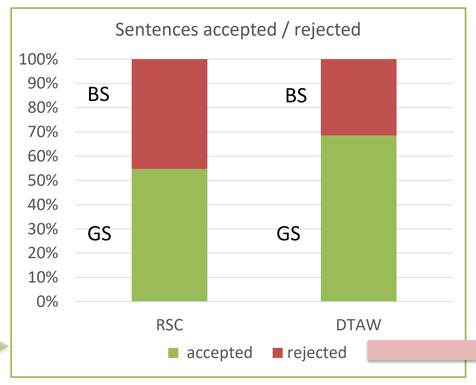
Case study and sanity check: Noun phrase development

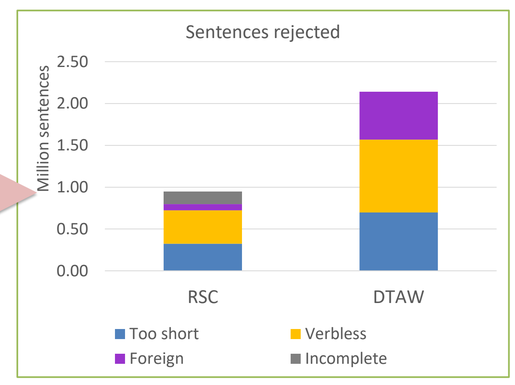
1. Does the data reflect previous observations? – Yes!
→ **MORE** NP internal modification with phrasal features – in the UD-framework: `nmod, appos, nummod, amod, det, compound, flat.`
→ **LESS** clausal postmodification, i.e., finite and non-finite clausal modifiers (`acl / acl:relcl`)

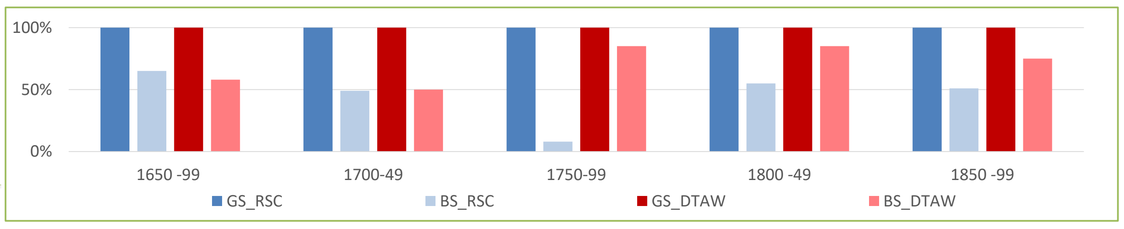
2. Is NP densification a cross-linguistic development? – Yes!
→ Significant decrease of clausal features in both EN and DE.

3. If so, is change driven by the same NP modifiers? – Almost!
→ EN and DE develop towards similar proportions of `amod` and `nmod`.

RESULTS

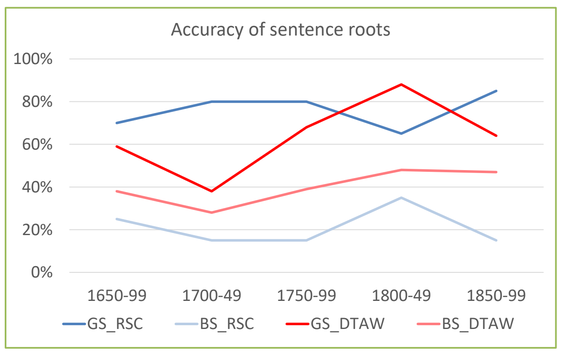


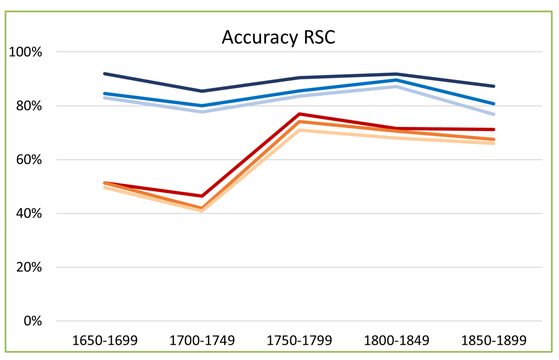


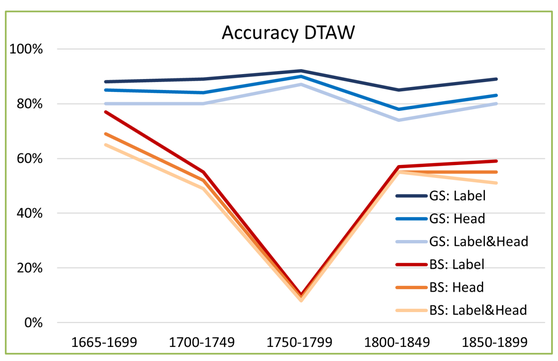


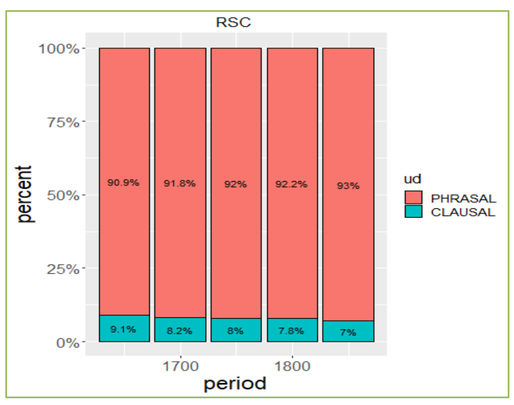
Period	RSC GS	RSC BS	DTAW GS	DTAW BS
1650–99	1	1.30	1.35	1.45
1700–49	1	1.30	2.50	1.45
1750–99	1	1.35	1.40	1.65
1800–49	1	1.05	1.20	1.35
1850–99	1	1.05	1.25	1.50
mean	**1**	**1.21**	**1.54**	**1.48**

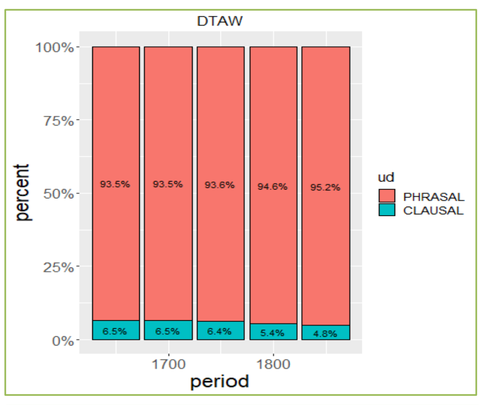
Table 3: Number of roots per sentence.

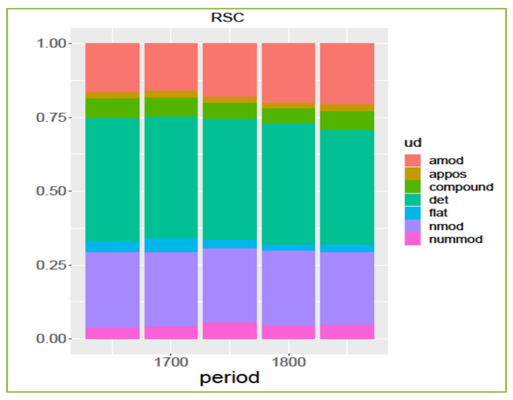


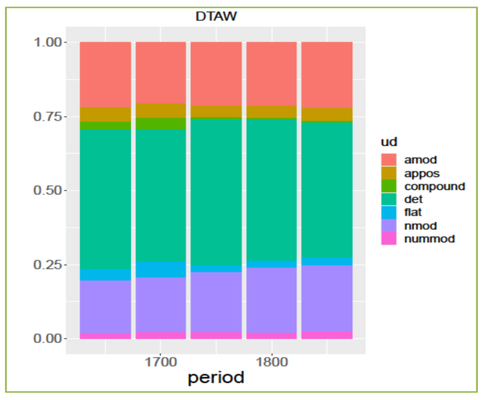












REFERENCES

[1] Fischer, S., J. Knappen, K. Menzel, and E. Teich. 2020. The Royal Society Corpus 6.0: Providing 300+ Years of Scientific Writing for Humanistic Study. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 794–802. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.99/

[2] Geyken, A., M. Boenig, S. Haaf, B. Jurish, C. Thomas, and F. Wiegand. 2018. 10. Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN. In *Digitale Infrastrukturen für die germanistische Forschung*, herausgegeben von Henning Lobin, Roman Schneider, und Andreas Witt, 219–48. De Gruyter. https://doi.org/doi:10.1515/9783110538663-011.

[3] Jurish, B. 2012. Finite-State Canonicalization Techniques for Historical German. Ph.D. thesis, Universität Potsdam, January.

[4] Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees. In Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK.

[5] Möslein, K. 1974. Einige Entwicklungstendenzen in der Syntax der wissenschaftlich-technischen Literatur seit dem Ende des 18. Jahrhunderts. *Zur Geschichte der deutschen Sprache und Literatur*, 94:156–198.

[6] Biber, D. and B. Gray. 2016. Grammatical complexity in academic English: Linguistic change in writing. *Studies in English Language*. Cambridge University Press.