Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4808–4816
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

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
Campus A2.2, 66123 Saarbrücken, Germany
{mariepauline.krielke, luigi.talamo}@uni-saarland.de
maib00001@stud.uni-saarland.de
j.knappen@mx.uni-saarland.de

Abstract

We present two comparable diachronic corpora of scientific English (RSC_UD-Parsed_1.0) and German (DTAW_UD-Parsed_1.0) from the Late Modern Period (17^{th}c.–19^{th}c.) annotated with Universal Dependencies. We describe several steps of data pre-processing and evaluate the resulting parsing accuracy showing how our pre-processing steps significantly improve output quality. As a sanity check for the representativity of our data, we conduct a case study comparing previously gained insights on grammatical change in the scientific genre with our data. Our results reflect the often reported trend of English scientific discourse towards heavy noun phrases and a simplification of the sentence structure (Halliday, 1988; Halliday and Martin, 1993; Biber and Gray, 2011; Biber and Gray, 2016). We also show that this trend applies to German scientific discourse as well. The presented corpora are valuable resources suitable for the contrastive analysis of syntactic diachronic change in the scientific genre between 1650 and 1900. The presented pre-processing procedures and their evaluations are applicable to other languages and can be useful for a variety of Natural Language Processing tasks such as syntactic parsing.

Keywords: universal dependencies, evaluation, English-German contrastive, diachronic linguistics, scientific language

1. Introduction

In the past years, interest in diachronic linguistic studies has grown. Thanks to new methods in corpus-based, computational diachronic studies such as diachronic word embeddings and deep neural networks to detect lexical semantic change, research in diachronic linguistics has advanced a great deal in understanding the evolution of the lexicon over time. However, efforts in unravelling grammatical change are less dynamic due to the lack of suitable and sound historical linguistic resources. For this reason, most previous studies on the diachronic syntactic development of scientific discourse in German (Möslein, 1974; Beneš, 1981; Habermann, 2011) and English (Halliday, 1988; Halliday and Martin, 1993) are purely descriptive, or, if corpus-based, limited to English (Biber and Gray, 2011; Biber and Gray, 2016).

In this paper, we present two Universal Dependency (UD) parsed corpora suitable to trace the syntactic development in the genre of scientific English and scientific German in the period between 1650 and 1900. This period, marked by the scientific revolution and a turn towards experimental science, is known as the beginning of modern science. To avoid pitfalls connected to the processing of historical language data, we propose several steps to make UD-parsing less error-prone. Familiar difficulties in working with historical data stretch from pre-processing to linguistic annotation. Due to not being digitally born, historical data require high pre-processing efforts including standardization of data formats, data cleanup (e.g. OCR errors) and meta-data derivation and annotation. Linguistic annotation bottlenecks are variation in spelling, morphology and syntax, e.g. in word order (Kermes et al., 2016; Menzel et al., 2021). As mentioned by Juzek et al. (2019), syntactic parsing most severely suffers from wrong sentence splitting. In this paper, we review previous approaches proposed for parsing of historical data (section 2), from which we select those suitable for our task of parsing historical English and German scientific texts. We present two historical, comparable resources of scientific language (English: RSC_UD-parsed_1.0 and German: DTAW_UD-parsed_1.0) covering the Late Modern Period (1650–1899) annotated with UD (section 3.1). We describe the corpora that we built as well as several pre-processing steps prior to the actual parsing to obtain the best possible parsing quality (section 3.2). We describe the parsing pipeline we specifically built for the two corpora, paying attention to the needs of the historical data and their cross-linguistic comparability (section 3.3). We then evaluate the resulting syntactic annotations and compare their accuracy to accuracy of non-preprocessed parses (see section 4). We show that pre-processing leads to a significant improvement of the parses. We then conduct a case study (section 5) focusing on the development of the noun phrase in scientific English and German with two aims. The first is to check for data sanity. Here, we assume that our English data should reflect trends in the development of the scientific genre previously mentioned in the literature. The second is to find out whether noun phrase densification is a cross-linguistic development occurring in both English and German. Our results confirm both assumptions. We conclude the paper with a summary (section 6) and provide an outlook for future research.

2. Related Work

In the following section, we will first give an overview over existing diachronic resources in both languages and standard ways of processing them. We then address previous studies on syntactic developments in the scientific genre in English and German.

2.1. Diachronic Resources and Available Tools

The number of comprehensive diachronic corpora that are freely available for scientific English and scientific German is extremely low. As for scientific English, available corpora are either specific to a particular field of science or relative to a certain time period (Kermes et al., 2016, 1928) and corpora mentioned therein, with the notable exception of the Royal Society Corpus (RSC: (Kermes et al., 2016). As for German, the scientific genre is covered only in multi-purpose diachronic corpora, such as the 'essay' genre of the TüBa-D/DC (Hinrichs and Zastrow, 2012), a diachronic corpus of German ranging from the 13thc. to the 20thc. and based on texts from the Gutenberg Project, or the several scientific disciplines in the Deutsches Textarchiv (DTA: (Geyken et al., 2018), a digital archive of texts from the 16thc. to the 19thc.

Such scarcity of resources is justified by the number and the complexity of tasks involved in the preparation and processing of historical data; for the purpose of the present work, we are concerned with two specific tasks: sentence splitting and dependency parsing.

As for sentence splitting, since historical texts do not have a consistent punctuation, the relatively easy task of punctuation disambiguation has to be replaced by the much harder task of sentence boundary detection (SBD), which was originally developed for transcribed text speech (Stevenson and Gaizauskas, 2000); however, SBD in historical data suffers from little research (Gerlof Bouma, 2013). An alternative approach to SBD is found in computational lexicography, where a common task is to automatically find in a corpus a "good sentence" to describe a lexical entry (Didakowski et al., 2012).

Dependency parsing is also problematic, since most of the historical languages are essentially low-resource languages; if the normalization task can improve other components of the NLP pipeline, such as the tokenization, the lemmatization and the PoS-tagging, this does not always apply to the dependency parsing (Juzek et al., 2019). Moreover, available dependency parsers are often trained on the news genre, thus performing worse with other genres. To the best of our knowledge, there are no comparative studies on the accuracy of the dependency parsing on the scientific genre (see however (Kanerva et al., 2020) on the parsing of biomedical texts); the best approximation is represented here by comparative studies taking into account the parsing of non-news genres; according to (Ortmann et al., 2019, 220), the two most accurate German dependency parsers for non-news genres are ParZu (Sennrich et al., 2009) and StanfordNLP (Qi et al., 2020); as for English, (Choi et al., 2015, 393) report Mate (Björkelund et al., 2010) and CoreNLP (Manning et al., 2014) as the NLP systems with the highest accuracy for "some genres" including the Bible and the Web.

A related issue, which largely remained unaddressed in the treatment of historical data, is the cross-linguistic adequacy of the annotations, especially with respect to the PoS tagging and the dependency parsing; the Universal Dependency (UD) project[1] (de Marneffe et al., 2021) offers a convincing and typologically-adequate framework to develop cross-linguistic annotations, currently featuring 200 treebanks for over 100 languages. Out of the four most accurate parsers mentioned above, only StanfordNLP is trained on a UD model, the UD German GSD (Ortmann et al., 2019, 216).

2.2. Syntactic Change in Scientific Language

Previous work on the diachronic development of scientific English (Halliday, 1988; Halliday and Martin, 1993; Biber and Gray, 2011; Biber and Gray, 2016) is strongly focused on the development of the noun phrase, consistently reporting a general trend from explicit verbal style towards heavy noun phrases leading to simplification of the overall sentence structure, e.g.(Atkinson, 1999; Banks, 2008). Studies on the syntactic development of scientific German describe a heavy influence by Latin syntax, resulting in dense sentence equivalent constructions, as well as a preference for intricate hypotactic structures over parataxis in the 17th and 18thc. In the 19thc. the trend reverses, sentences become shorter, subordination less frequent while nominalizations increase (Möslein, 1974; Beneš, 1981). Contrastive studies on scientific English and German are scarce, however Krielke (2021) shows that both languages decrease the use of relative clauses over time.

3. Data and Methods

In the following section, we present our two scientific corpora (RSC and DTAW). We describe several pre-processing steps prior to parsing and modifications made to the parsing pipeline.

3.1. Corpora

Royal Society Corpus (RSC) For scientific English, we use the Royal Society Corpus (Kermes et al., 2016). The corpus covers almost 250 years of scientific texts taken from the *Philosophical Transactions* and *Proceedings* of the Royal Society of London between 1665 until 1899 (Fischer et al., 2020). The original version contains ca. 32 million tokens with standard linguistic annotation. Normalization of historical word forms

[1] https://universaldependencies.org

was implemented using VARD (Baron and Rayson, 2008), tokenization, lemmatization and part-of-speech tagging was created with TreeTagger (Schmid, 1994).

Deutsches Textarchiv Wissenschaft (DTAW) The German corpus features scientific texts from the Deutsches Textarchiv (DTA, (Geyken et al., 2018)) between 1650 and 1899. The corpus size of this portion of the corpus is ca. 82 million tokens before applying our pre-processing. The DTA comes with canonicalized word forms created with CAB (Jurish, 2012). Tokenization of the DTA is done using the specifically built tool DTA-Tokwrap (Jurish, 2012).

Since both corpora already feature valuable linguistic annotation based on customized processing, we decide to maintain as many of the annotations as possible to facilitate the parsing process. Based on previous annotations, we employ further pre-processing steps tailored to the specific technical requirements of each corpus. Corpus data were prepared as follows.

3.2. Pre-processing

Normalization of Historical Data. We additionally normalize the DTAW corpus for punctuation replacing the formerly common virgule (slash) by the analogous comma (Example (1)).

(1) *Wann jemand etwas seinem Nächsten zum Besten aufrichtig heraus gibt / so gering es auch ist / billig zu Dank soll angenommen werden. (DTAW, Glauber, Opera Chymica, 1658)*

Extraction of "good sentences". To minimize the number of end-of-sentence-errors, we apply several rules to extract "good sentences" (**GS**) only. For this, we build on preexisting annotation to detect non-sentential constructions as well as foreign-language sentences. Specifically, we deleted sentences beginning with a word in lower case and the sentence preceding them (*incomplete*), sentences with less than 8 tokens (*too short*), as well as sentences lacking a verb (*verbless*). To exclude foreign-language sentences, we ran the language recognizer LangID (Lui and Baldwin, 2011) on each sentence in the two corpora and excluded all sentences in other languages than the language of the corpus. After pre-processing, we obtain approximately 26 million tokens for the English corpus and 74 million tokens for the German corpus. For detailed information of accepted tokens and sentences after applying the above rules see table 1. For a comparative evaluation of the improvement gained by the **GS** selection, we also retained all discarded "bad sentences" (**BS**). The German corpus was processed with version 2.0.0 of the script that did not yet implement *incomplete*.

3.3. UD-parsing

The texts are extracted from the two pre-processed corpora in such a way that metadata are preserved. Before parsing the texts with UDpipe 1 (Straka and Straková, 2017), we preserve the original sentence splitting and tokenization. As the name suggests, the parser uses models from the Universal Dependencies project (de Marneffe et al., 2021): GSD for German and GUM for English. Both models are trained on multi-genre data including academic texts (GUM) and encyclopedic articles (GSD). We believe these two models to be a good fit for our data since, first there are no models exclusively trained on scientific texts, and second, because the scientific genre was in its very early stages at the beginning of our observed time period. Thus, older texts still show more general language features presumably covered by a multi-genre model.

Since the German UD-tagset does not include the `acl:relcl` tag to identify relative clauses we furthermore enrich the German treebank with this information by applying the following rule: any token tagged as `acl` with a child whose POS tag is PRELS (substituting relative pronoun) or PRELAT (attributive relative pronoun) should be renamed as `acl:relcl`.

The corpora resulting from pre-processing and enriched with UD-parses are then called RSC_UD-Parsed_1.0 (English) and German DTAW_UD-Parsed_1.0 (German), for the sake of space, in this paper, we will refer to them as **RSC** and **DTAW**.

3.4. Code Availability

The script for the extraction of "good sentences" is available on github here,[2] the other scripts for the above mentioned improvements are available on github here.[3]

4. Parser Evaluation

To evaluate the quality of the parses after the pre-processing steps described above, we sample 100 sentences (20 from each 50 years period, e.g. 1650–1699) from the "good sentences" (**GS**) and evaluate them against 100 parsed sentences from those discarded by our filter (**BS**). The samples are evaluated by linguistic experts according to three different aspects: parsability of a sentence, number and accuracy of roots, and parsing accuracy itself.

Parsability. We evaluate if the parser can be expected to make sense of a sentence, i.e., if the sentence shows any kind of grammatically interpretable structure for a particular language. We accept title-like noun phrases (see Ex. (2)) as well as dates (see Ex. (3)), but we exclude sentences in other languages than English or German (see Ex. (4)) and fragments without grammatical, linguistically parsable structure such as equations (see Ex. (5)) and accumulations of abbreviations (see Ex. (6)).

(2) *Section of a villus, from the small intestine of a monkey.*

[2] `https://github.com/SFB1102/B1-gute-saetze`

[3] `https://github.com/MhmudFwzi/corpora_analysis`

	RSC		DTAW	
	#	%	#	%
Tokens				
processed	31,952,725	100.00	82,461,237	100.00
accepted	**26,127,595**	**81.77**	**74,692,952**	**90.58**
rejected	5,825,130	18.23	7,768,285	9.42
Sentences				
processed	1,119,141	100.00	3,127,793	100.00
accepted	**612,458**	**54.73**	**2,142,839**	**68.51**
rejected	506,683	45.27	984,954	31.49
Too short	323,935		697,071	
Verbless	399,488		870,503	
Foreign	74,197		573,350	
Incomplete	146,639		—	

Table 1: Corpus size in terms of accepted tokens and sentences after pre-processing. A sentence can be rejected for more than one reason therefore the numbers in the last four rows don't sum up to the total of rejected sentences.

(3) *Feb. 4, 1800.*

(4) *Explication de la Feuille de Landen.*

(5) *r I.23+ I.6.9 n8 r.-1195 n.=8 Log. 28.9= 1.46090 8.*

(6) *deg. , and Latitude 34.*

Our results (see Table 2) show that for both languages (RSC, English; DTAW, German) the selection for "good sentences" (**GS**) was 100% successful, i.e., all of the retained sentences are parsable. The numbers for parsability of a "bad sentence" (**BS**) show that in English more sentences that are actually parsable were discarded, while in the texts from newer periods fewer of the bad sentences were parsable. This is due to a higher number of equations in the newer data on the one hand, and a higher number of sentences consisting of noun phrases in the older data on the other hand. For German, we find an opposite trend: our pre-processing excluded more actually parsable sentences from newer data than in the older data. This is due to a much higher number of foreign sentences in the older data, while the "bad sentences" from newer time periods include a high number of defective sentence splittings (*incomplete*) resulting in sentence fragments which are still syntactically interpretable. All resulting parsability values for "bad sentences" are significantly below the values obtained for "good sentences".

	RSC		DTAW	
Period	GS	BS	GS	BS
1650 – 1699	1.00	0.65	1.00	0.58
1700 – 1749	1.00	0.49	1.00	0.50
1750 – 1799	1.00	0.08	1.00	0.85
1800 – 1849	1.00	0.55	1.00	0.85
1850 – 1899	1.00	0.51	1.00	0.75
mean	**1.00**	**0.46**	**1.00**	**0.71**

Table 2: Evaluation of parsability of a sentence

Roots. We evaluate the number and accuracy of roots per sentence. A well-parsed sentence should only have one root. We check how many roots are assigned to one sentence and evaluate if the assignment is correct. We find that for English, UDpipe consistently assigns exactly one root to each of the **GS**, while assigning more than one root to the **BS** (Table 3). Also, accuracy is significantly higher (t = 10.126, df = 7.943, p-value = 8.138e-06) for the **GS** than for the **BS** (see Table 4). For German, root detection does not seem to work very well, neither for the **GS** nor for **BS** (see Table 3). Average numbers of roots per sentence in **GS** and **BS** do not vary significantly (t = 0.24244, df = 4.3257, p-value = 0.8195), which shows that the processing does not improve a one-root-per-sentence only processing of the German parser. The detection of several roots per sentence in German therefore rather seems to be due to parser-internal issues. However, the accuracy of root detection (Table 4) is significantly better for the **GS** than for the **BS** (t = 2.7498, df = 5.8555, p-value = 0.03415).

	RSC		DTAW	
Period	GS	BS	GS	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.

UD-annotation. Following the example of Spacy's accuracy evaluation[4], we evaluate correctness of the assigned UD-label (Label) per token (cf. DEP_LAS in Spacy's evaluation scheme), correctness of the syntactic head (Head) of each token (cf. DEP_UAS in

[4] `https://spacy.io/models/de`

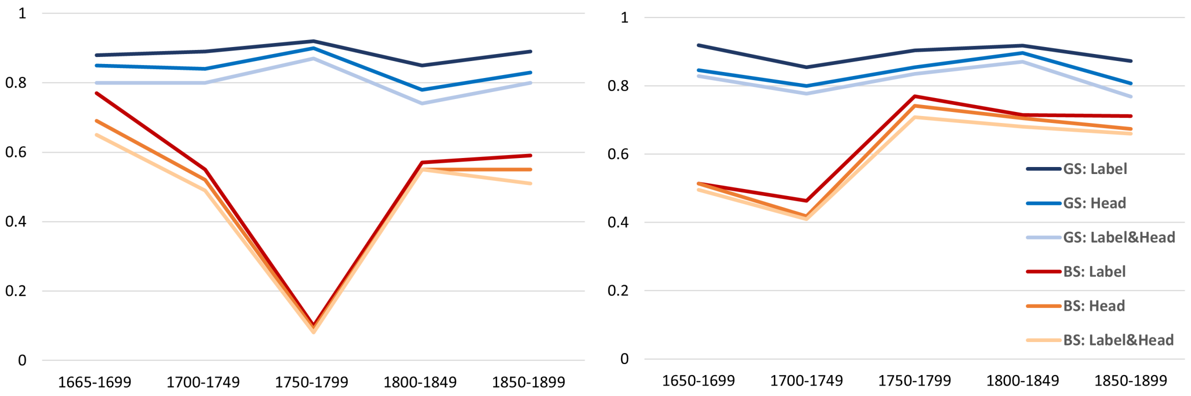

Figure 1: Accuracy of UD Label and Head in **RSC** (left) and **DTAW** (right) by 50-year periods.

	RSC		**DTAW**	
Period	GS	BS	GS	BS
1650–99	0.70	0.25	0.59	0.38
1700–49	0.80	0.15	0.38	0.28
1750–99	0.80	0.15	0.68	0.39
1800–49	0.65	0.35	0.88	0.48
1850–99	0.85	0.15	0.64	0.47
mean	**0.76**	**0.21**	**0.63**	**0.40**

Table 4: Accuracy of roots.

	Label		**Head**		**Label&Head**	
Period	GS	BS	GS	BS	GS	BS
1665–99	0.88	0.77	0.85	0.69	0.80	0.65
1700–49	0.89	0.55	0.84	0.52	0.80	0.49
1750–99	0.92	0.10	0.90	0.09	0.87	0.08
1800–49	0.85	0.57	0.78	0.55	0.74	0.55
1850–99	0.89	0.59	0.83	0.55	0.80	0.51
mean	**0.88**	**0.52**	**0.84**	**0.48**	**0.80**	**0.46**

Table 5: **RSC** Evaluation of parses of good sentences (GS) vs. bad sentences (BS) : correct UD-tags, correct recognition of syntactic head, correct UD-tag and head.

	Label		**Head**		**Label&Head**	
Period	GS	BS	GS	BS	GS	BS
1650–99	0.92	0.51	0.85	0.51	0.83	0.50
1700–49	0.85	0.46	0.80	0.42	0.78	0.41
1750–99	0.90	0.77	0.85	0.74	0.84	0.71
1800–49	0.92	0.72	0.90	0.71	0.87	0.68
1850–99	0.87	0.63	0.81	0.67	0.77	0.66
mean	**0.89**	**0.65**	**0.84**	**0.61**	**0.82**	**0.59**

Table 6: **DTAW** Evaluation of parses of good sentences (GS) vs. bad sentences (BS) : correct UD-label, correct recognition of syntactic head, correct UD-label and head.

Spacy's evaluation scheme), and correctness of both labels (Label and Head) per token. Accuracy is calculated as the number of correctly annotated tokens over the whole number of tokens in a time period. We conduct evaluations for **GS** as well as **BS**. The parse of a non-parsable sentence is regarded as entirely incorrect, since for such a sentence no actual correct parse exists. Figure 1 shows that for both languages the **GS** have a much higher accuracy on all levels (Label and Head) than the **BS**. Across all time periods and in both languages, the accuracy values for **GS** differ significantly ($p < 0.05$) from **BS** showing that our pre-processing improves parsing accuracy significantly. For **English** (Table 5), accuracy of "good sentences" is constantly near 90% for Label and near 80% for correct detection of the syntactic head (Head). On average, both UD-label and head were assigned correctly in 80% of the evaluated **GS** tokens. We do not find an accuracy improvement over time, in fact, t-tests for all time periods compared to each other show no significant difference for the accuracy values encountered for each period. Looking at the English **BS**, we see that parsing quality drops towards the end of the 18^{th}c. and increases afterwards (Figure 1). The extremely low accuracy derives from the low number of actually parsable sentences in the time period 1750–1799. A look into the **BS** reveals an abundance of abbreviations (e.g., *Exp. los!.*) and equations (*n-1 X 1/≈≈1.*), reducing parsability.

For **German** GS, we find slightly higher accuracy for Label and Head than for the English data (see Table 6) with values between 80 and 90%. As well as for English, the GS values do not differ significantly from each other according to time period, which shows that parsing quality of "good sentences" does not improve significantly with more modern data. This suggests that our pre-processing contributes to a stable parsing quality throughout the observed time periods.

Note that for both languages the Head accuracy is always lower than the Label accuracy. This could be due to the parser's performance itself. However, it is also possible that annotators have a general tendency to accept a UD-label as correct, since the task is more difficult than determining the correct syntactic head.

Overall, our evaluations have shown that the employed pre-processing steps help improve parsing quality significantly on all three levels: parsability, root accuracy and UD-annotation (label and head detection). For English, our pre-processing also contributes significantly to preventing parses from containing more than one root.

5. UD-Analysis

In the following analysis, we focus on noun phrase modification features previously described as becoming distinctive for scientific English (Biber and Gray 2016) and check whether our English corpus reflects the described tendencies. We then compare the observed development to our German corpus. We start by inspecting the development of the noun phrase over time splitting the corpus in 50 years periods. We distinguish between UD-relations representing phrasal features, which contribute to complexity within the noun phrase and clausal features, which modify the noun phrase by clausal subordination. Phrasal features create rather implicit relations between modifiers and their head noun, while clausal features are grammatically explicit (Biber and Gray, 2016) specifying at least for subject and verb. Phrasal features in the UD-framework are nominal dependents, such as nouns as nominal dependents of another noun (`nmod`), appositional modifiers (`appos`), numeric modifiers (`nummod`), adjectival modifiers (`amod`) and determiners (`det`), as well as multiword expressions such as compounds (`compound`) and composite names (`flat`). Clausal features are finite and non-finite clausal modifiers of a noun (`acl` / `acl:relcl`). Our analysis shows that in line with previous findings for scientific English (especially (Biber and Gray, 2011; Biber and Gray, 2016)) the phrasal features gradually become (significantly) more frequent over time making the noun phrase more complex on a phrasal level and moving away from complex clausal subordination (see Figure 2)

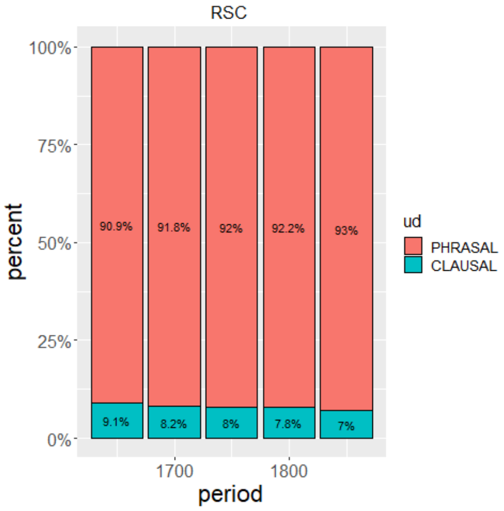

Figure 2: Development of phrasal and clausal modifiers in scientific English (RSC).

For German, we find a similar trend. In line with qualitative studies (Möslein, 1974; Beneš, 1981), the decline of the clausal features and the increase in phrasal features is time shifted towards the end of the 19th c. The differences between the distributions of nominal and clausal features are non-significant between 1650 and 1700 but significant between all later periods (Figure 3).

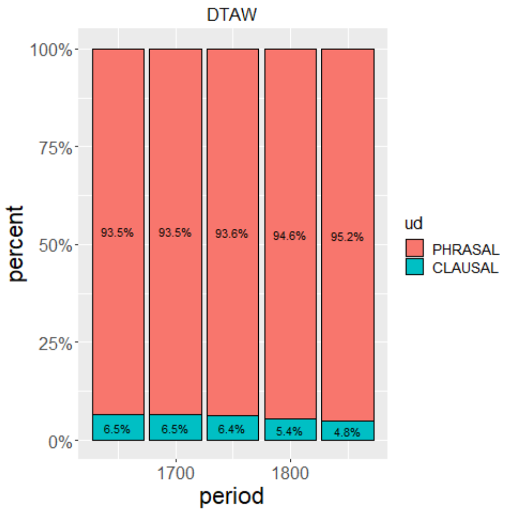

Figure 3: Development of phrasal and clausal modifiers in scientific German (DTAW).

To see whether in both languages change is driven by the same noun phrase modifiers, we take a closer look at the phrasal features (Figure 4). For English, we find a notable increase in attributive adjectives (`amod`) reflecting findings by Biber and Gray (2016). Also, numeric modifiers (`nummod`) take over an increasing proportion, which is plausible since scientific discourse in the wake of the scientific revolution increasingly becomes based on numbers. However, composite names (e.g., *Sir Isaac Newton*) become less frequent. Determiners take the highest proportion of all noun phrase modifiers in both languages but are fairly stable over time.

In German, the most striking relative increase within the phrasal modifiers can be found for nominal NP modifiers (e.g., *Ende des fünften Teils*, `nmod`). This finding is interesting, since it highlights an opposite trend in noun phrase modification in the two languages. While English starts out at a rather low proportion of adjectival modifiers (`amod`), German shows a stable use of adjectival modifiers with a proportion of approximately 20%, which is the same proportion English reaches in 1850 after a gradual increase over time. Conversely, German starts out at a relatively low proportion of nominal post-modifiers (`nmod`), gradually climbing up to the diachronically stable proportion found for scientific English (approx. 25%). As a result, at the end of the 19th c., English and German show strikingly similar proportions of adjectival pre-modification and nominal post-modification. For nu-

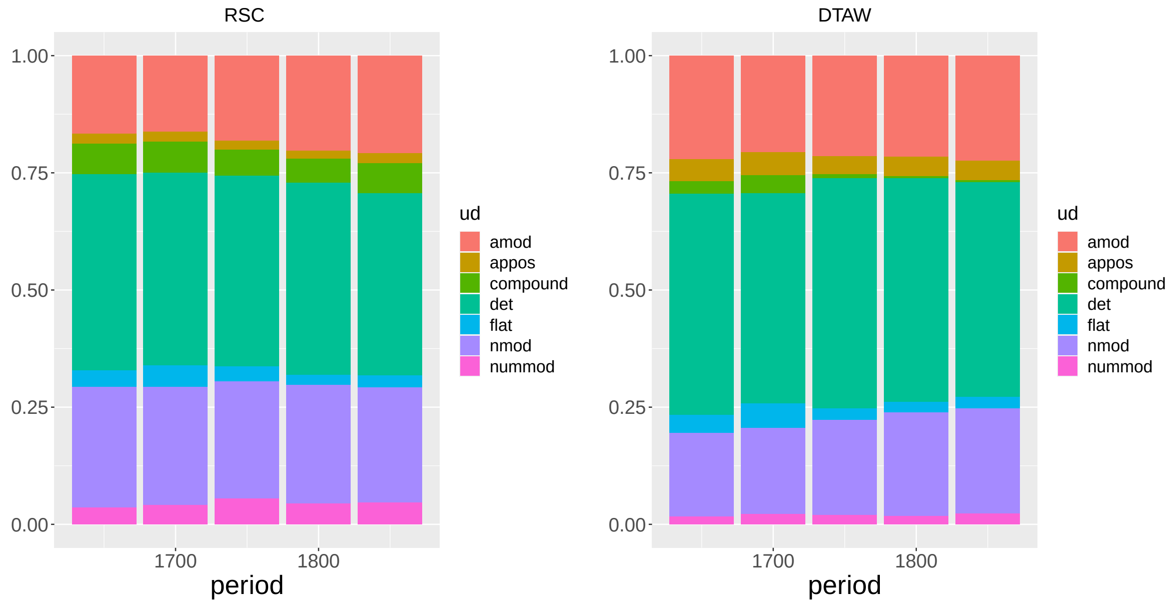

Figure 4: Development of **phrasal** NP modifiers (fine grained) in **RSC** (left) and **DTAW** (right) by 50-year periods.

meric modifiers the development for German is similar to that in English (slight increase) as well as that of composite names (slight decrease). We also find a steep decrease of compounds in German, especially from 1700 to 1750, which is due to orthographic conventionalization of German compounds as one word (e.g., *Baum-Früchte – Baumfrüchte*). The remaining compounds in the later periods are mostly combinations of two adjectives (e.g., *romanisch-germanisch*) or split off parts of compounds as in *König- und Kaisertum*.

The findings of this analysis reflect the claim that scientific language becomes less explicit over time. According to Biber and Gray (2016) (chapter 6.4) explicitness is created by clausal post-modification, while phrasal constituents such as attributive adjectives and nominal post-modifiers (ibid.) create rather non-explicit relations between the head noun and its modifier. Both scientific corpora show a trend away from clausal post-modification and an increasing reliance on adjectival pre-modifiers as well as nominal post-modifiers. Thus, our analysis has shown to be a valid sanity check for the parsing quality of the English corpus, since the results coincide with the relevant literature (Biber and Gray, 2011; Biber and Gray, 2016; Halliday, 1988; Halliday and Martin, 1993). For German, the analysis has given us a first idea that the development on the level of the noun phrase is similar to that of English scientific language. We also found that both languages become increasingly similar with respect to noun phrase modification.

6. Conclusion

We have presented two comparable, diachronic corpora of scientific English (RSC_UD-Parsed_1.0) and German (DTAW_UD-Parsed_1.0) annotated with Universal Dependencies (UD). We described several preprocessing steps to prepare historical data for UD parsing. By evaluating the parses with and without preprocessing, we showed that these steps significantly improve parsing quality and help achieve an accuracy of >80% for both historical corpora. In a case study focusing on constituents in noun phrases, we further checked parsing quality by comparing the encountered trend to existing findings. We found that our data for English confirm the reported trend towards less clausal and more phrasal structures in a noun phrase. We compared these findings to German showing the same trend. We further discovered that English and German noun phrases pattern in increasingly similar ways over time, strongly relying on adjectival (pre-)modifiers, determiners and nominal (post-)modifiers. The corpora are available via the Saarland University CLARIN-D centre[5] under a FAIR license and can be used for syntactic analyses of the scientific genre in the Late Modern period. In future versions, the German corpus will be recompiled with exclusion of *incomplete* sentences and both corpora will be augmented with dependency length and depth to trace cross-linguistic diachronic shifts towards shorter dependencies in scientific language as done for English by Juzek et al. (2020).

7. Acknowledgements

This work is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

[5] 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`

8. Bibliographical References

Atkinson, D. (1999). Language and science. *Annual Review of Applied Linguistics*, 19:193–214.

Banks, D. (2008). *The development of scientific writing. Linguistic features and historical context*. Equinox, London, Oakville.

Beneš, E. (1981). Die formale Struktur der wissenschaftlichen Fachsprachen aus syntaktischer Hinsicht. In Theo Bungarten, editor, *Wissenschaftssprache*, pages 185–212. Fink, München.

Biber, D. and Gray, B. (2011). The historical shift of scientific academic prose in English towards less explicit styles of expression. *Researching specialized languages*, 47:11.

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

Choi, J. D., Tetreault, J., and Stent, A. (2015). It depends: Dependency parser comparison using a web-based evaluation tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 387–396, Beijing, China, July. Association for Computational Linguistics.

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

Didakowski, J., Lemnitzer, L., and Geyken, A. (2012). Automatic example sentence extraction for a contemporary German dictionary. In Ruth Vatvedt Fjeld et al., editors, *Proceedings of the 15th EURALEX International Congress*, pages 343–349, Oslo,Norway, aug. Department of Linguistics and Scandinavian Studies, University of Oslo.

Fischer, S., Menzel, K., Knappen, J., and Teich, E. (2020). The Royal Society Corpus 6.0 providing 300+ years of scientific writing for humanistic study. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*. ELRA.

Gerlof Bouma, Y. A. (2013). Experiments on sentence segmentation in Old Swedish editions. In *Proceedings of the 19th Nordic Conference of Computational Linguistics, NODALIDA 2013, May 22-24, 2013, Oslo University, Norway*, volume 87 of *Linköping Electronic Conference Proceedings*, pages 11–26. Linköping University Electronic Press.

Habermann, M. (2011). *Deutsche Fachtexte der Neuzeit. Naturkundlich-medizinische Wissensvermittlung im Spannungsfeld von Latein und Volkssprache*. De Gruyter, Berlin/ Boston.

Halliday, M. A. K. and Martin, J. R. (1993). *Writing science: Literacy and discursive power*. Falmer Press, London.

Halliday, M. A. K. (1988). On the language of physical science. In Mohsen Ghadessy, editor, *Registers of written English: Situational factors and linguistic features*, pages 162–177. Pinter, London.

Juzek, T., Fischer, S., Krielke, M.-P., Degaetano-Ortlieb, S., and Teich, E. (2019). Annotation quality assessment and error correction in diachronic corpora: Combining pattern-based and machine learning approaches. In *52nd Annual Meeting of the Societas Linguistica Europaea*.

Juzek, T. S., Krielke, M.-P., and Teich, E. (2020). Exploring diachronic syntactic shifts with dependency length: The case of scientific English. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 109–119.

Kanerva, J., Ginter, F., and Pyysalo, S. (2020). Dependency parsing of biomedical text with BERT. *BMC Bioinformatics*, 21, December.

Krielke, M.-P. (2021). Relativizers as markers of grammatical complexity: A diachronic, cross-register study of English and German. *Bergen Language and Linguistics Studies*, 11(1):91–120.

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.

Ortmann, K., Roussel, A., and Dipper, S. (2019). Evaluating off-the-shelf NLP tools for German. In *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pages 212–222, Erlangen, Germany.

Stevenson, M. and Gaizauskas, R. (2000). Experiments on sentence boundary detection. In *Sixth Applied Natural Language Processing Conference*, pages 84–89, Seattle, Washington, USA, April. Association for Computational Linguistics.

9. Language Resource References

Baron, A. and Rayson, P. (2008). VARD 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Birmingham.

Björkelund, A., Bohnet, B., Hafdell, L., and Nugues, P. (2010). A high-performance syntactic and semantic dependency parser. In *Coling 2010: Demonstrations*, pages 33–36, Beijing, China, August. Coling 2010 Organizing Committee.

Geyken, A., Boenig, M., Haaf, S., Jurish, B., Thomas, C., and Wiegand, F. (2018). 10. Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN. In Henning Lobin, et al., editors, *Digitale Infrastrukturen für die germanistische Forschung*, pages 219–248. De Gruyter.

Hinrichs, E. and Zastrow, T. (2012). Automatic annotation and manual evaluation of the diachronic German corpus TüBa-D/DC. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1622–

1627, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Jurish, B. (2012). *Finite-State Canonicalization Techniques for Historical German*. Ph.D. thesis, Universität Potsdam, January. (completed 2011, published 2012).

Kermes, H., Degaetano-Ortlieb, S., Khamis, A., Knappen, J., and Teich, E. (2016). The Royal Society Corpus: From Uncharted Data to Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1928–1931, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Lui, M. and Baldwin, T. (2011). Cross-domain feature selection for language identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 553–561, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Menzel, K., Knappen, J., and Teich, E. (2021). Generating linguistically relevant metadata for the Royal Society Corpus. *Research in Corpus Linguistics, Challenges in combining structured and unstructured data in corpus development (special issue)*, 9(1):1–18.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

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.

Sennrich, R., Schneider, G., Volk, M., and Warin, M. (2009). A new hybrid dependency parser for German. In *Proceedings of the GSCL Conference, Potsdam, Germany*.

Straka, M. and Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.