Relativizers as markers of grammatical complexity: A diachronic, cross-register study of English and German[1]

Marie-Pauline Krielke

University of Saarland (Germany)

In this paper, we investigate grammatical complexity as a register feature of scientific English and German. Specifically, we carry out a diachronic comparison between general and scientific discourse in the two languages from the 17th to the 19th century, using relativizers as proxies for grammatical complexity. We ground our study in register theory (Halliday and Hasan, 1985), assuming that language use reflects contextual factors, which contribute to the formation of registers (Quirk *et al*., 1985; Biber *et al*., 1999; Teich *et al*., 2016). Our findings show a clear tendency towards grammatical simplification in scientific discourse in both languages with English spearheading the trend early on and German following later.

Keywords: contrastive linguistics, corpus linguistics, diachronic linguistics, English/German

1. Introduction

In the present paper, we look at the period between 1650 and 1900, which is especially interesting, since academic disciplines and with them scientific discourse emerges (Görlach, 2004). The development of new expressive structures reflects new communicative needs (cf. Betten, 2016). Register theory assumes that different text classes not only differ from general language in topic or field, but also in terms of lexico-grammatical features reflecting tenor and mode. This has been shown in numerous corpus-linguistic studies (Biber, 1988, 1993, 2006, 2012). Teich *et al.* (2016) follow the hypothesis that the development of scientific language undergoes two parallel processes, specialization and diversification. They show that, over time, scientific communication becomes increasingly expert-oriented, and the different scientific disciplines develop their own distinct set-ups of lexico-grammatical features. Specifically, for scientific English, previous research has shown a clear development towards higher lexical density (Biber, 2006; Aarts *et al*., 2012; Biber and Gray, 2016; Degaetano-Ortlieb *et al.*, 2016) alongside a simplification in syntax (Halliday, 1988; Teich *et al.*, 2016). German syntax, however, is described as becoming increasingly complex during the 17th and 18th century due to a strong remaining Latin influence and only in later periods, a trend towards detangling this

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

CROSSING THE BORDERS: ANALYSING COMPLEX CONTRASTIVE DATA. Edited by Anna Čermáková, Signe Oksefjell Ebeling, Magnus Levin and Jenny Ström Herold. *BeLLS* Vol 11, No 1 (2021), DOI: 10.15845/bells.v11i1.3440.

complex syntax is observed (Möslein, 1974; Beneš, 1981; Admoni, 1990; Habermann, 2011). Based on the findings for the two languages, we assume that decreasing grammatical complexity may be a cross-lingual register feature shaping scientific discourse over time.

We investigate grammatical complexity on the example of full, finite relative clauses (RCs). Being clausal postmodifiers, RCs represent the most explicit and syntactically most intricate way of defining a referent (as compared to alternative structures such as attributive adjectives and prepositional phrases) since at least a subject (specifying the agent) and a verb (marked for tense, aspect and mode) are included, as illustrated in example (1a). Transformations of RCs to less explicit structures, illustrated by the postmodifying prepositional phrase in example (1b), lead to processing difficulties characteristic of scientific language including (among others) lexical density, syntactic ambiguity and grammatical metaphor (cf. Halliday, 1988).

(1)

a) He affirms, that he has been the First *that* has discovered that Vessel, *which* by him is called Salivare Exterius. (Philosophical Transactions, 1665−1678)

b) He affirms his discovery *of the Salivare Exterius Vessel*. (generated alternative)

Due to clausal embedding example (1a) is grammatically more complex than example (1b). Accumulations of RCs within one sentence represent especially strong cases of syntactic complexity, as illustrated in example (2).

(2) Next, that the two Eyes were united into one Double Eye, *which* was placed just in the middle of the Brow, the Nose being wanting, *which* should have separated them, *whereby* the two Eye-holes in the Scull were united into one very large round hole, into the midst of *which*, from the Brain, entered one pretty large Optic Nerve, at the end of *which* grew a great Double Eye; that is, that Membrane, called Sclerotis, *which* contained both, was one and the same, but seemed to have a Seam, by *which* they were joined, to go quite round it, and the fore or pellucid part was distinctly separated into two Cornea's by a white Seam *that* divided them. (Philosophical Transactions, 1665−1678)

Besides the most common relativizers *which* and *that*, pronominal adverbs are another, highly explicit way of conveying a relationship between the antecedent and the subject of the RC, as in example (3a).

(3)

a) […] the Membrane immediately encompassing that skin, *wherein* the Faetus is wrapped […]. (Philosophical Transactions, 1665−1678)

b) […] the Membrane immediately encompassing that skin wrapping the Faetus […]. (generated alternative)

As seen in example (3a), the prepositional specification of location *in* (in where*in*) is implicitly entailed in the verb *wrap* and could therefore be omitted (3b). Omission of any kind of superfluous grammatical information could be a counterbalance to other sources of informational overload, such as continuously new emerging vocabulary. Besides lower syntactic intricacy, a reduced set of alternatives at a given choice point, i.e., fewer different relativizers and better predictability of the specific options reduce grammatical complexity through reduction of entropy.

2. Related work

Relative clauses (RCs) are a widely studied topic in English diachronic as well as synchronic linguistic studies (Ball, 1996; Nevalainen and Raumolin-Brunberg, 2002; Hundt *et al.*, 2012; Nevalainen, 2012), in vernacular varieties of English (Romaine, 1980, 1982; Tottie and Harvie, 2000; Tagliamonte, 2002; Tagliamonte *et al.*, 2005; Levey, 2006) and in spoken and written mode (Guy and Bayley, 1995; Lehmann, 2001). Diachronic as well as synchronic studies (Biber *et al.*, 1999; Leech *et al.*, 2009; Hinrichs *et al.*, 2015) on relativizer choice find that the selection of relativizers largely depends on the overall formality level of a text, *which* being the formal option whereas *that* becomes increasingly common to informal text types. However, there are only few studies reflecting on the use of pronominal adverbs in relativizer position. Mellinkoff (2004), for instance, mentions their diachronic integration in the language of the law, and Österman (1997) points to their primary association with formal genres. Diachronically, Krielke *et al.* (2019) have shown a remarkable decrease in pronominal adverbs in scientific English between 1650 and 1850. This paradigm reduction of pronominal adverbs over time can partly be explained by the typological drift from synthetic to analytic (Nevalainen and Raumolin-Brunberg, 2012), e.g., *whereby* becoming *by which*.

RCs and their and alternative syntactic renderings (prepositional phrases and attributive adjectives), as well, have received ample scholarly attention for their role as frequent constituents in noun phrases. For written discourse, Biber *et al.* (1988) report on a strong preference for (premodified) nouns and postmodifying prepositional phrases, while spoken registers rather rely on embedded clauses. Biber and Finegan (1997) show a steady trend towards nominal structures in the past 300 years of academic writing. Biber and Gray (2011) specifically mention a slight decrease in RCs in academic texts, again pointing to a remarkable increase in phrasal as compared to clausal modification creating a compressed academic style in present day English (see e.g., Halliday, 1988; Biber and Clark, 2002; Mair, 2006; Biber and Conrad, 2009).

The aforementioned studies largely rely on patterns of parts-of-speech, but also syntax-based studies found a decrease in RCs as compared to nominal premodifications (see for instance Juzek *et al.*, 2020). The predominantly frequency-based approaches, however, ignore ambient context. Information-theoretic measures, such as surprisal and entropy, considering the probabilities of linguistic units given their syntagmatic contexts have proven to be important factors driving linguistic change (Degaetano-Ortlieb and Teich, 2016; 2019; Rubino *et al.*, 2016). Especially convergence on specific grammatical features, which can be measured by conditioned probabilities, over time leads to conventionalization, a mechanism giving way to innovation on the lexical level. Degaetano-Ortlieb and Teich (2019) give a unified explanation of the evolution of the scientific register based on the assumption that register evolution depends on evolving communicative needs while striving for the creation of an optimal code customized for communication between experts. This code is assumed to be characterized by specific linguistic features to balance information load. For our study we adopt Teich *et al.*'s (2016) assumptions regarding register shifts formulated in their hypotheses on specialization: The development of a scientific field leads to increasing expert-orientation. Expert-orientation manifests itself along two dimensions: a) increasing technicality and information density, linguistically expressed by nominal style and high lexical density, and b) decreasing grammatical intricacy of the sentence structure, i.e., the number of clauses in the sentence and their interdependencies (cf. Halliday, 1988; Halliday and Martin, 1993). To measure grammatical intricacy, Teich *et al.* (2016) inspect, amongst other features, the number of clauses (including RCs) as well as the number of relativizers per sentence in scientific texts between the 1970s and the 2000s. For our study this points to the assumption that RCs are a

feature of scientific discourse, however intricacy in the sense of embeddedness of many clauses within one sentence is rather specific to general language.

In contrast to the studies of English, studies of German on diachronic grammatical change are more qualitative in nature. The observed time period in this study starts at the end of the third and last period of Early New High German, a time period bringing forth a variety of new, especially informative text types promoting increasing distinctiveness between general language and the language of the learned (1550−1700; Admoni, 1990).

Habermann's (2001) comprehensive account on the development of German syntax in the natural sciences between the 15th and 19th century focuses on the influence of Latin on the emerging vernacular German as a language of scientific communication. Scientists received their education in Latin, influencing their lexical as well as syntactic style. Preferred structures influenced by Latin were, for instance, sentence equivalent short forms pursuing information density, while expanding hypotaxis with deep embeddings was preferred over parataxis.

Möslein (1974) describes the syntactic developments in scientific-technical literature since the end of the 18th century. Due to the establishment of verb final position in the 17th century (starting in technical literature), main and subclause can be distinguished from each other. Formation of long and embedded sentences to present complex thoughts in one sentence becomes possible leading to an extreme increase in hypotactic structures in the 17th and 18th century (ibid.). Starting in the first half of the 19th century, a trend of disentanglement and reduction in sentence length as well as a remarkable reduction in subordinate clauses and an increase in nominalizations is described to take place in scholarly German (Möslein, 1974; Beneš, 1981). Societal developments of the time, such as increasing influence of mass media and other European languages of science, are reflected in a new trend towards lower syntactic intricacy. As a result, scientific style became increasingly condensed aiming for clarity and efficiency of expression in response to evolving communicative needs. Possible reasons for the increase of nominal groups instead of subclauses could be exactness and effort reduction (see for instance *dependency locality theory*, Gibson *et al.*, 2000). Factors that may lead to reduction in cognitive effort are compound formation as an alternative to prepositional phrases (*iron oxide vs. oxide of iron*), and nominalizations instead of subclauses avoiding grammatical complexity connected to tense, mode and number (Möslein, 1974).

The studies on the different German relativizers we are aware of (Ebert, 1986; Reichmann and Wegera, 1993; von Polenz, 1999; Ágel, 2000; Fleischer, 2004; Brooks, 2006; Dal, 2014; Pickl, 2020) only look at the standard relativizers, *der/die/das (d-)* being the most frequent relativizer and *welcher/welche/welches (welch-)* being the marked, formal variant (see Pickl, 2020) in isolation, while a comprehensive view on relativizers, including pronominal adverbs, is still lacking. Mentioned as promoters of syntactic intricacy (Möslein, 1974, Admoni, 1990), RCs have also been analyzed with information-theoretic measures. Voigtmann and Speyer (2020), for instance, use surprisal (Shannon, 1949; Levy, 2008; see section 4.2) to detect information density related preference for RC extraposition, assuming that extraposition is used as a strategy to counterbalance an informational overload in the RC and spread information evenly across a sentence. Krielke *et al.* (2019) use surprisal to detect increasingly predictable contexts of the relativizer *which* in English, finding that it is increasingly used in adverbial gaps (see Biber *et al.*, 1999), particularly to express relations of manner (*the means by which, the manner in which*). Further, Krielke *et al.* (2019) use entropy (see section 4.2) to trace paradigmatic variety of a group of relativizers including prepositional adverbs similar to Milin *et al.* (2009), who measure entropy over inflectional paradigms. In the present study we use entropy to measure paradigmatic variability of the relativizer paradigm, as well as surprisal to account for the predictability of relativizers in German and English over time. We assume that in scientific discourse the contexts of RCs become increasingly predictable over time, thus contributing to the overall processing ease of otherwise complex concepts.

3. Hypotheses

We use relativizers as proxies to investigate the development of grammatical complexity for English and German, pursuing the following hypotheses:

1. In the course of register formation, scientific discourse becomes grammatically less complex in terms of

   a) *syntactic intricacy*, as indicated by the frequency of relativizers and the number of RC embeddings within a sentence, decreases as register formation evolves.

   b) *paradigmatic richness*, i.e., the available types of relativizers, decreases over time as indexed by entropy.

   c) *contextual predictability* of relativizers, i.e., relativizers appear in increasingly similar contexts as indexed by surprisal.

Since German is described to initially expand its syntactic possibilities during the 18th century and due to much later institutionalization of scientific discourse in German, we expect the development to manifest along different trajectories, leading to the second hypothesis:

2. English should show a more linear development, while we expect German to first increase syntactic intricacy and paradigmatic richness before decreasing towards the 19th century.

4. Data and Methods

4.1 Data

For scientific English (SE), we use the *Royal Society Corpus* (RSC v4.0; Kermes *et al.*, 2016), consisting of the *Proceedings and Transactions of the Royal Society of London* covering the time from 1665−1869 with approximately 32 million tokens. For general English (GE), we use the *Corpus of Late Modern English Texts* (CLMET; Diller *et al.*, 2011), spanning 1710−1920 with approximately 40 million tokens from several genres (e.g., narrative, drama). For German, texts from 1650−1900 are retrieved from the scientific (SG) and general language (GG) subcorpora of *Deutsches Textarchiv* (DTA, Geyken *et al.*, 2018) respectively. Scientific German is represented with approximately 80 million tokens, general German with approximately 60 million tokens including non-fictional as well as fictional prose texts. All subcorpora contain metadata (e.g., author, publication year) and linguistic annotation (e.g., tokens, lemmas, normalization, parts-of-speech, surprisal). Part-of-speech annotation for English is based on the "Penn Treebank Tagset" (Santorini, 1990), for German on the "Stuttgart-Tübingen Tagset" (STTS, Thielen *et al.*, 1999).

4.2 Methods

To trace the development of grammatical complexity in the scientific register in the two languages, we focus on the features shown in Table 1. We apply conventional frequency-based methods to account for syntactic intricacy indicated by the frequency of relativizers in the four subcorpora, as well as the number of RC embeddings within a sentence.

Table 1. Features of grammatical complexity.

Discourse Property	Feature Category	Feature subcategory	Measure
Grammatical complexity	Grammatical Intricacy	Frequency of relativizers Relativizers per sentence	Relative frequencies
	Paradigmatic richness	Relativizer paradigm (number of different relativizers)	Entropy (H)
	Contextual predictability	Probability of relativizers given their context	Surprisal

To account for predictability of items in context, we use information-theoretic measures such as surprisal (as operationalized by Degaetano-Ortlieb *et al.*, 2016) of the different relativizers. To assess paradigmatic changes (growth or reduction) in the group of available relativizers, we calculate entropy. Finally, we qualitatively investigate the top three preceding trigrams sequences (part-of-speech and lexical) representing highly predictive contexts of relativizers.

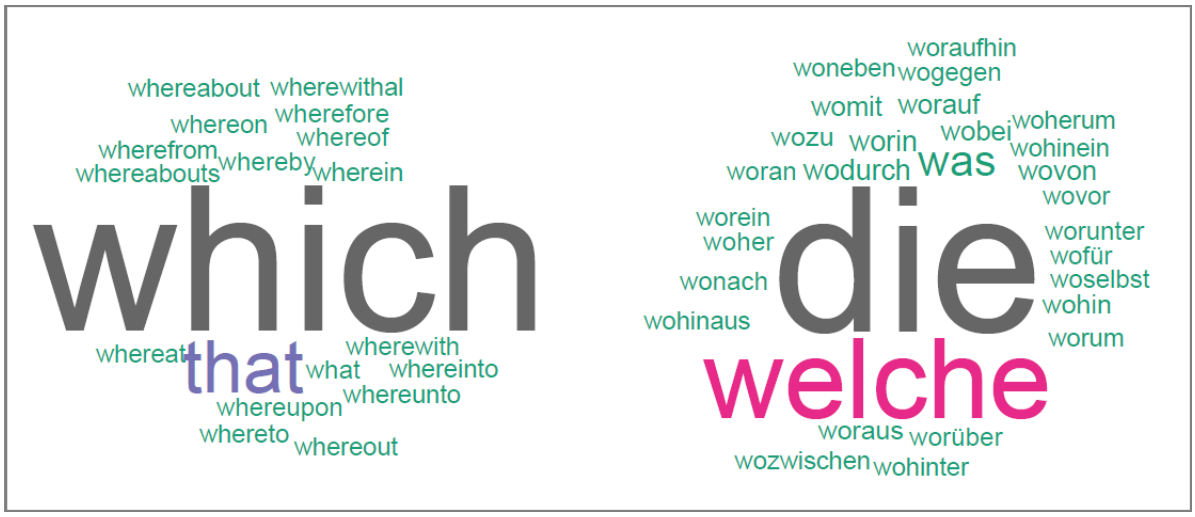

Figure 1. Relativizer paradigms in English (left) and German (right), color and size indicate frequency.

To grasp the full historical extent of the paradigms (apart from the most common ones *which* and *that* and *welcher, welche, welches (welch-)* and *der, die, das (d-)*), we first determine the existing members of the paradigms. For English, we extract all words beginning with *where-* and sort out all words not representing pronominal adverbs and, for German, all words beginning with *wo(r)-* and part-of-speech (POS) tagged as PRELS/PRELAT, resulting in the lists provided in figure 1. The motivation to use information-theoretic measures is the assumption that language users strive for effort reduction on the one hand and successful communication on the other. Previous studies have shown that production effort is directly linked to the number of options at a given choice point (Milin *et al.*, 2009). Fewer encoding options lead to entropy reduction (cf. hypothesis 1b). For the relativizer paradigm we can assume that fewer available relativizers lead to lower production effort for the sender of a message, as well as lower comprehension effort for the receiver of the message, since that receiver will have a more confined expectation and lower uncertainty of the upcoming word. We use entropy to measure the uncertainty about a set of choices at a given point. Formally, entropy is the expected (weighted average) amount of information in a paradigm. The more members the paradigm has and the more similar the probabilities of the different members are, the higher the entropy. Thus, entropy is highest if all probabilities are equal. We calculate the entropy of the English and German relativizer paradigms (figure 1) to find whether there is a register specific trend for entropy reduction and if so, whether this is the case in both languages.

Register specific preference and with it a reduction in paradigmatic entropy should lead to convergence on conventionalized linguistic choices.

Entropy is directly related to surprisal, i.e., the negative *log* probability of a word to occur in a certain context (Crocker *et al.*, 2015). The higher the probability of a word in a particular context, the less surprising is its occurrence in this context (Degaetano-Ortlieb *et al.*, 2016). We calculate surprisal based on the conditional negative *log* probabilities from a 4-gram language model, i.e., the negative log probability of a word given its three preceding words. For our analysis, we are interested in the distributions of the surprisal values of the observed three groups of relativizers (*which, that, (welch-), (d-)* and pronominal adverbs). To visualize the surprisal distributions of each relativizer group, we use boxplots displaying the distribution of the individual surprisal values indicating five different measures: minimum, first quartile, median, third quartile and maximum. The boxplots also show whether the values are symmetrically distributed, how tightly the values are grouped and if they are skewed. We assume that, over time, relativizers will occur in increasingly predictable contexts (have a lower surprisal), ensuring successful and effortless communication (cf. hypothesis 1c).

5. Analysis

5.1 Syntactic intricacy

We aim to trace changes in syntactic intricacy via two measures. First, we calculate (a) relative frequencies of the whole group of relativizers per subcorpus per 50 years, assuming that a higher number of relativizers represents a higher number of RCs and therefore a preference for post-modification. Second, we calculate (b) the average number of relativizers per sentence per 50 years, assuming that a higher number of relativizers per sentence represents a stronger tendency towards embeddedness. (a) is displayed in figure 2 (for English) and 3 (for German).

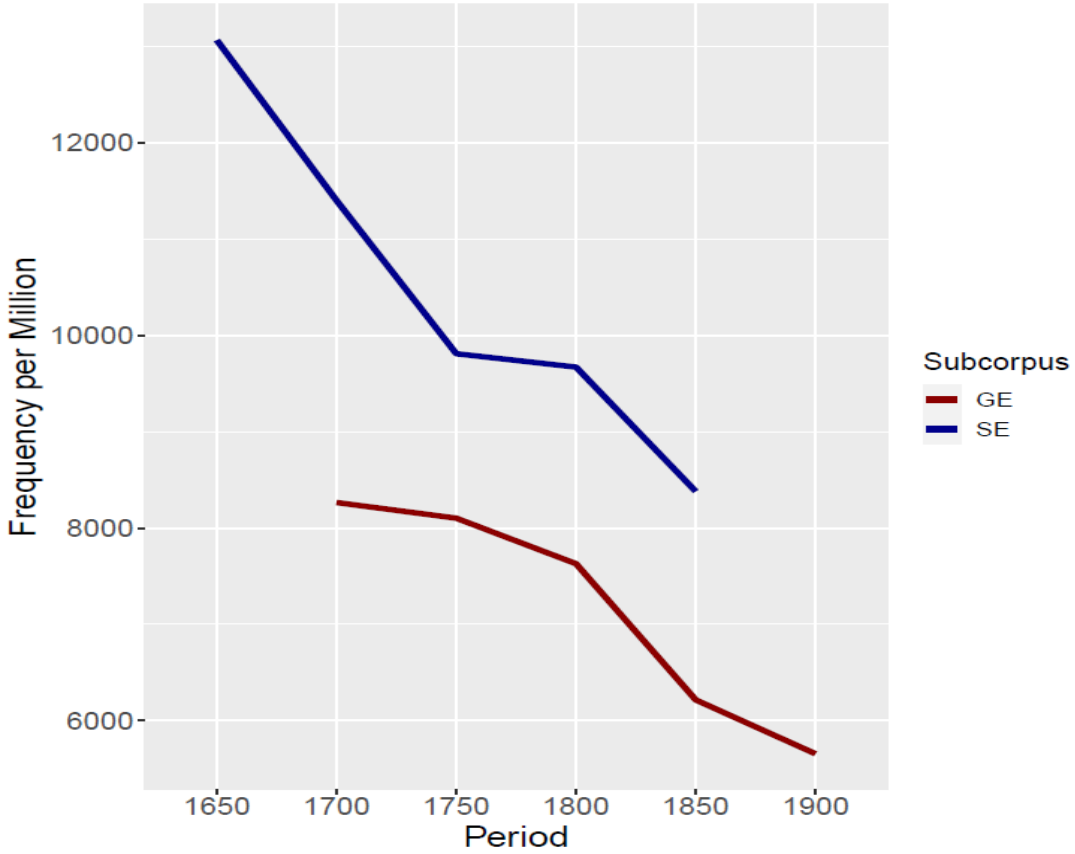

Figure 2. Frequency per million of relativizers in general (GE) vs. scientific English (SE).

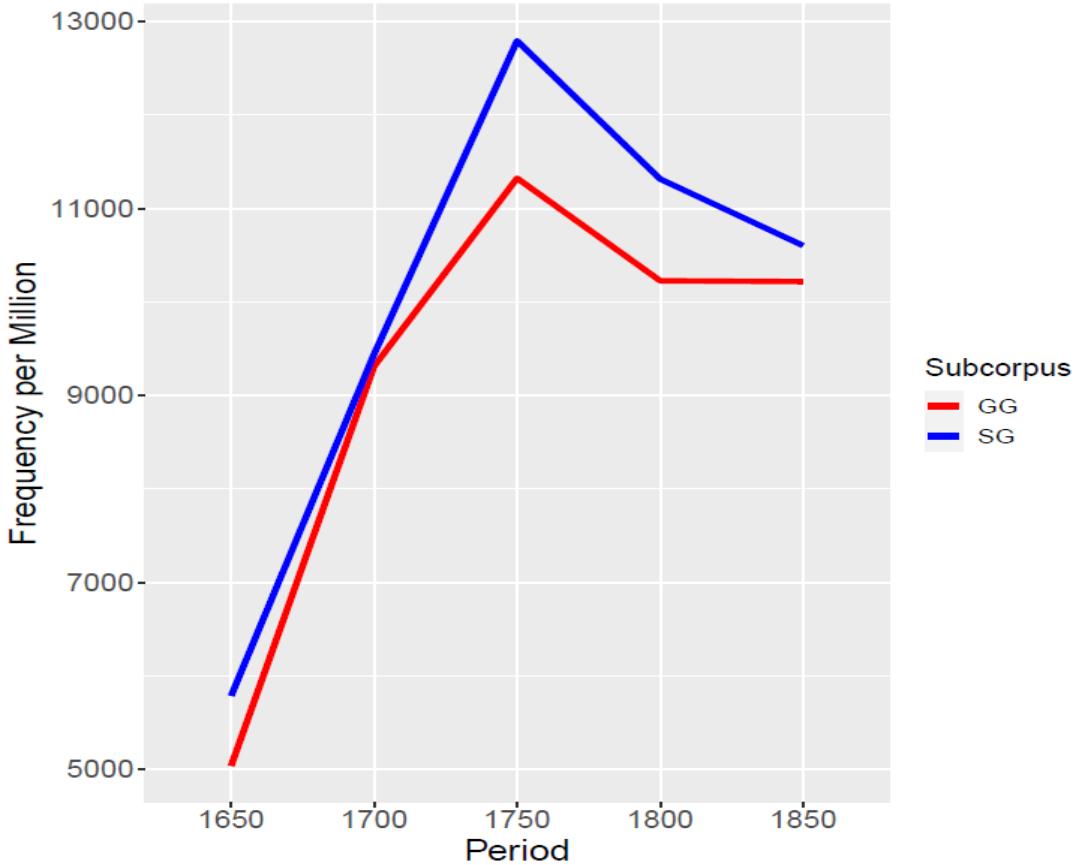

Figure 3. Frequency per million of relativizers in general (GG) vs. scientific German (SG).

First, between languages, we see strongly diverging trends, while within languages trends are quite similar. For English, we see a steady decrease of relativizers in both scientific and general language. In the scientific texts, relative frequencies start out almost twice as high and with a much steeper downward trend than in the GE texts. This shows an overall preference for using RCs in SE in the earlier time periods and a clear development towards a less embedded syntax over time. In German, the frequencies in both subcorpora are quite similar until 1750 and start to diverge afterwards. Also, throughout the observed time period SG shows higher frequencies of relativizers than GG. Frequencies peak in 1750 followed by a decrease in both SG and GG. The peak in SG, however, is more pronounced. In the first half of the 19th century frequencies stabilize in GG, while further declining in SG. In contrast to English, German starts out at a much lower frequency of relativizers between 1600 and 1650 only reaching the starting frequency of SE in 1750, indicating that in German hypotaxis was expanding throughout the baroque period, as also suggested by Habermann (2001) and Admoni (1990). The trends differ between the two languages until 1750 and align afterwards indicating that in German the trend towards simpler syntax became popular only in the 18th century as suggested by Admoni (1990). This is comprehensible considering the strong influence of Latin stylistic ideals German scientific text production was under.

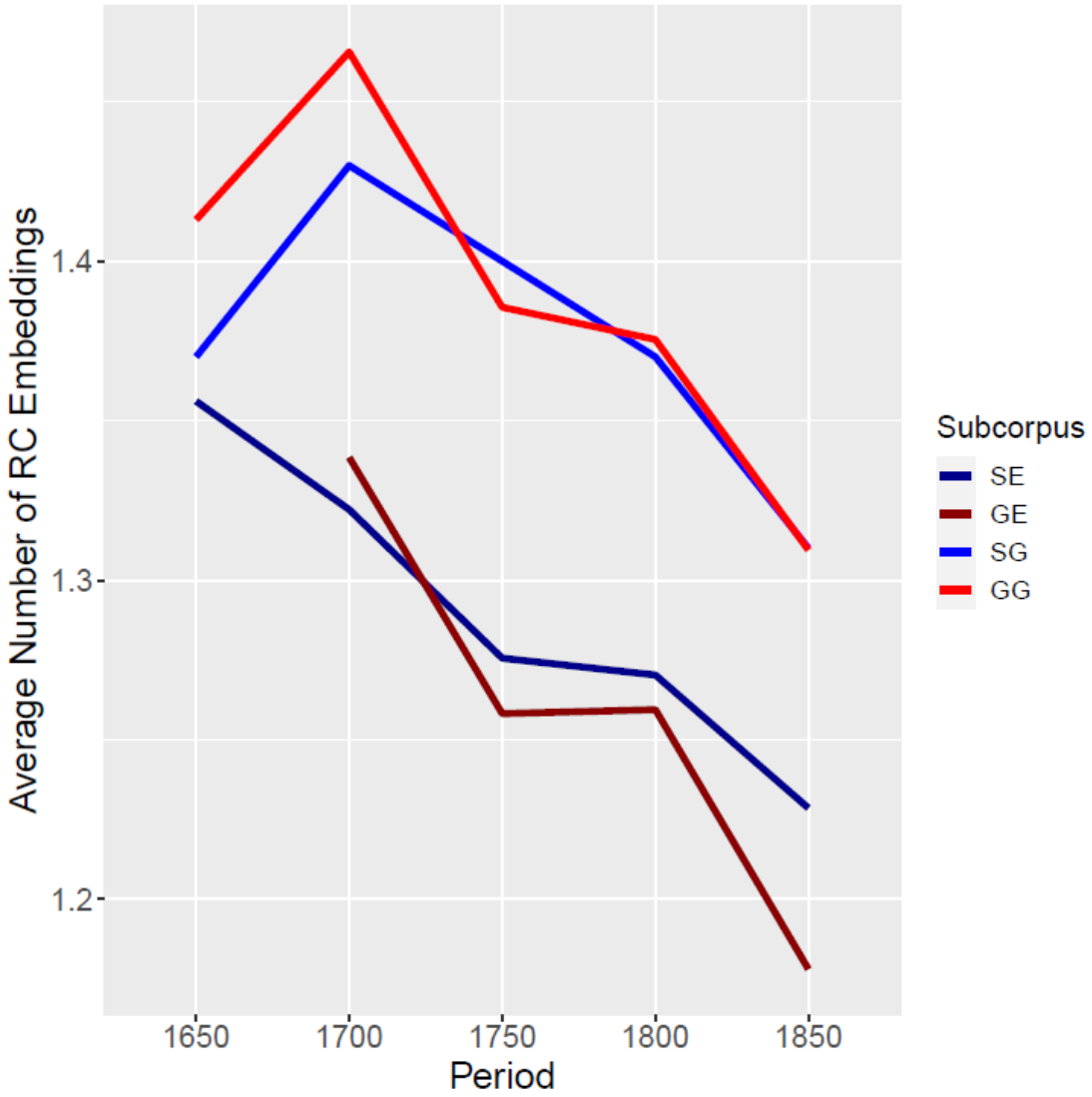

Figure 4. Average number of RCs per sentence in English (GE & SE) and German (GG & SG).

Looking at the average number of RCs embedded in a sentence (figure 4), we find that the trends broadly coincide with the shapes of the frequency distributions. In the first half of the 18th century embeddedness is stronger in GE than in SE, while there are overall more relativizers used in SE than in GE. This indicates that GE overall made use of fewer RCs, which, however, often occurred within one sentence. In the second half of the 18th century the trend reverses. Scientific English shows stronger embeddedness together with a higher number of RCs overall. In both SG and GG, RC embeddings show a steep increase towards the first half of the 18th century and an equally steep decrease afterwards representing the flourishment of clause embeddings in the 17th and 18th century. Interestingly, SG overall shows fewer embeddings per sentence than GG (with an exception between 1750 and 1800) while constantly showing a higher frequency in relativizers. This points to a need to employ explicit structures to explain complex matters while not overstretching the boundaries of cognitive processing load. Another interesting fact is that RC embeddedness peaks earlier than the overall frequency of relativizers. This points to a rather unbalanced use of relativizers in the first half of the 18th century: fewer relativizers overall clustering together in fewer sentences. In the second half of the 18th century this trend reverses: more relativizers overall are spread across different sentences.

5.2 Paradigmatic richness

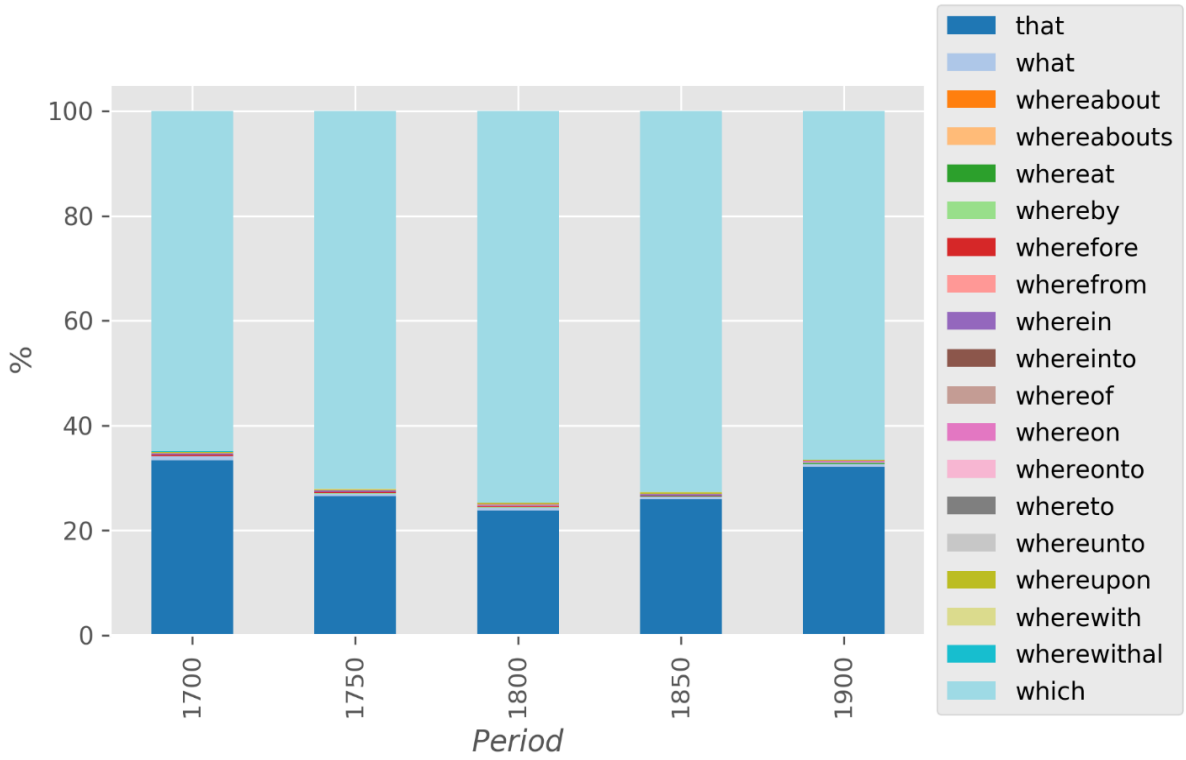

Figure 5. Distribution of different relativizers in GE.

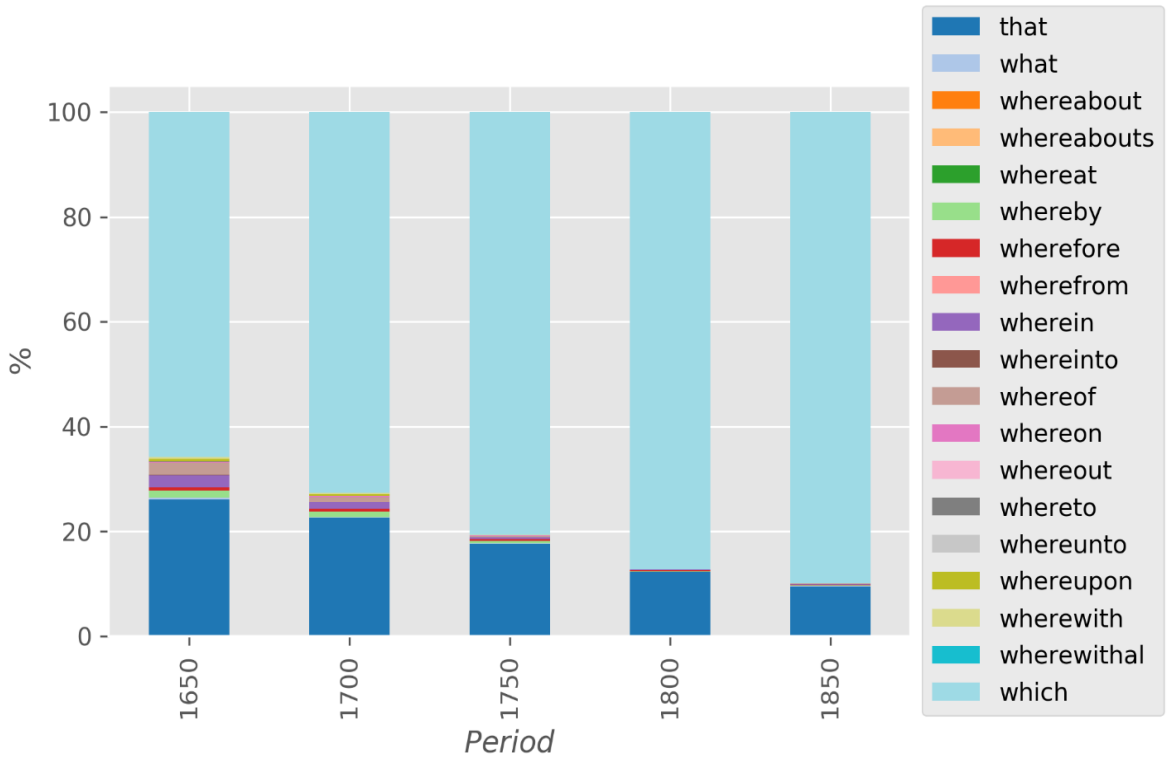

Figure 6. Distribution of different relativizers in SE.

The English subcorpora differ substantially regarding relativizer distribution, making relativizers a clear register feature in the early periods. While GE (figure 5) shows a stable distribution of the different relativizers with *which* being the overall most frequent type, SE (figure 6) starts out with a great variability of available relativizers including a large group of pronominal adverbs. For the GE subcorpus, we see that pronominal adverbs throughout all time periods occupy a rather negligible proportion. Looking at the distribution of the different relativizer types in SE, we find the biggest variability of relativizers in 1650. Together with overall decreasing relative frequencies of relativizers, variability gradually decreases, too. Over time, *which* becomes increasingly dominant, pushing out all other relativizers to under 10% in 1850. The gradual decrease of pronominal adverbs is in line with observations by Nevalainen and Raumolin-Brunberg (2012) and Krielke *et al.* (2019), confirming the abandonment of synthetic forms. In addition, this outcome confirms our intuition about differentiation of the scientific register against the general language by converging on a preferred linguistic feature and substituting a variety of alternatives. This is in line with observations of conventionalization in the scientific domain by Degaetano-Ortlieb and Teich (2019) and Teich *et al.* (2021). We will show this even more clearly by calculating entropy of the relativizer paradigms in each subcorpus.

In German, we find an inverse picture. Figure 7 shows that, like SE, in GG pronominal adverbs become less frequent. In SG (figure 8), in contrast, pronominal adverbs take up an increasing portion of the paradigm until 1850 and abruptly decrease after 1850. In exchange, *(welche-)* becomes more frequent taking the place left by the pronominal adverbs in decline. This is interesting for two reasons. First, German academic style seems to prefer a diverse set of options to introduce RCs in an explicit way. Only towards the end of the 19th century both frequency and relativizer variety seem to decline, possibly following the example of other European scientific traditions as suggested by Möslein (1974) and Beneš (1981). Second, while overall relativizer frequency was already on the wane, productivity of relativizers was still in

expansion: This points to an even greater variability of the paradigm in the period between 1800 and 1850.

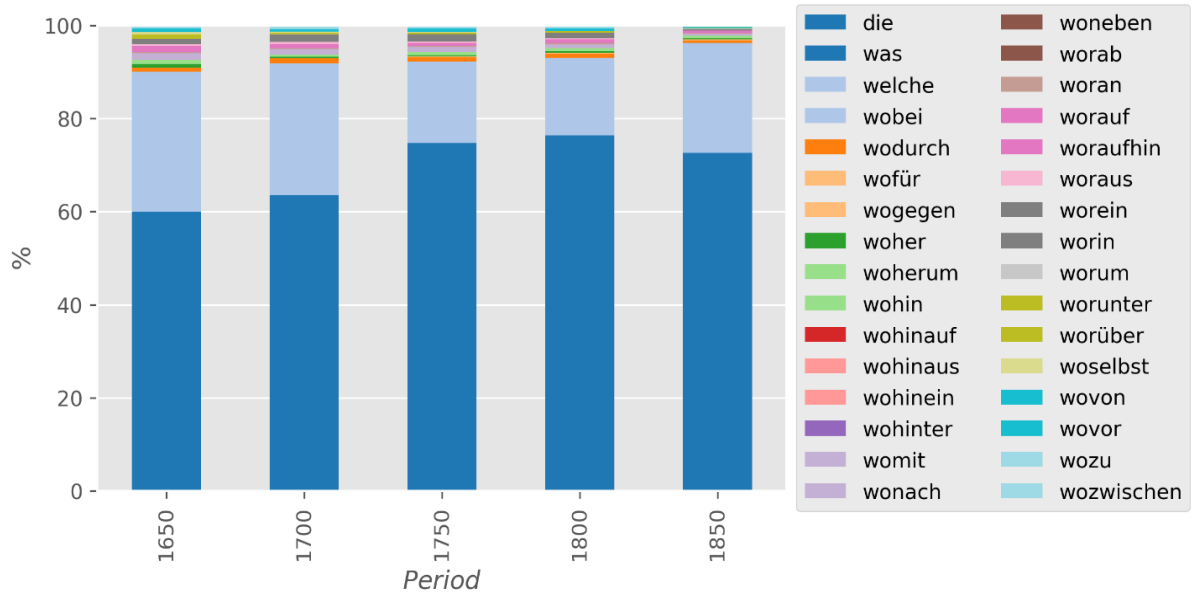

Figure 7. Distribution of different relativizers in GG.

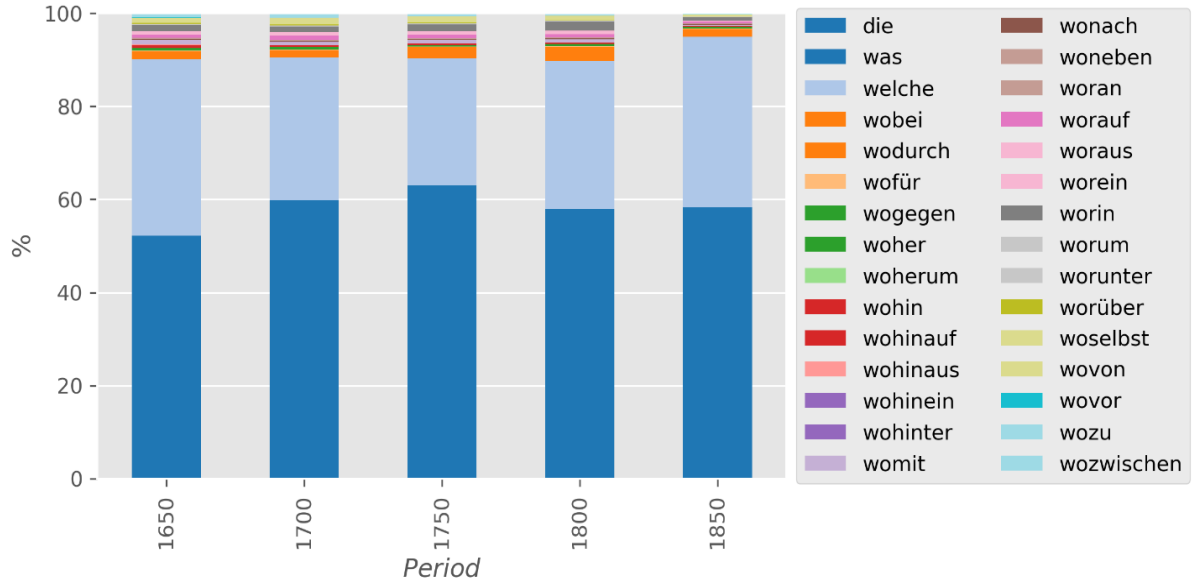

Figure 8. Distribution of different relativizers in SG.

Looking at entropy, we find relatively stable values in GE (figure 9), while for SE (figure 10) we see a striking reduction of entropy over time. The entropy trends clearly reflect the distributional trends in figures 5–6, while also considering predictability: the reduction in entropy in SE over time is owed to a smaller choice of options between the different relativizers on the one hand, but also to an increasing probability of *which* to occur as compared to decreasing probabilities of all other available options.

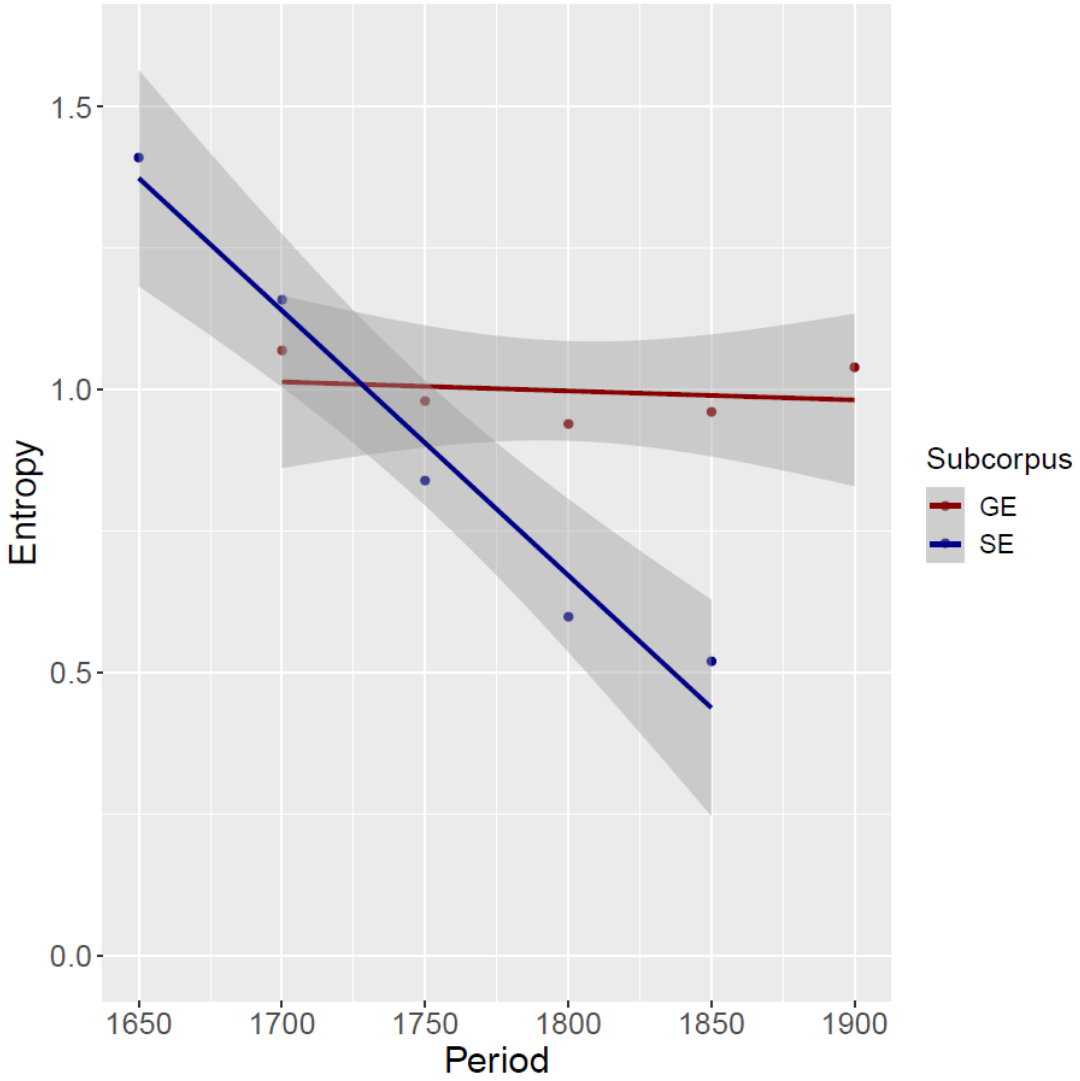

Figure 9. Entropy over the relativizer paradigm in general (GE) and scientific English (SE).

Figure 6 shows that in 1650 scientific writers had a much bigger choice amongst different relativizers than in 1850. At the same time, readers of scientific texts in 1650 had a much higher uncertainty about the upcoming relativizer than a reader in 1850. In GE (figure 5), the choice/uncertainty did not change over time. The entropy value of 1 points at a choice between two preferred options, presumably *that* and *which*. The entropy value in SE in 1850 is around 0.5, a third of the value in 1650, indicating a strong preference for *which* as the relativizer of choice. This again reflects the tendency towards conventionalization of options, as observed by Teich *et al.* (2021). In German, the paradigm of relativizers is much bigger than in English (compare figure 1), contributing to overall higher entropy values (German ranges between 1.5 and 2, while English ranges between 0.5 and 1.5).

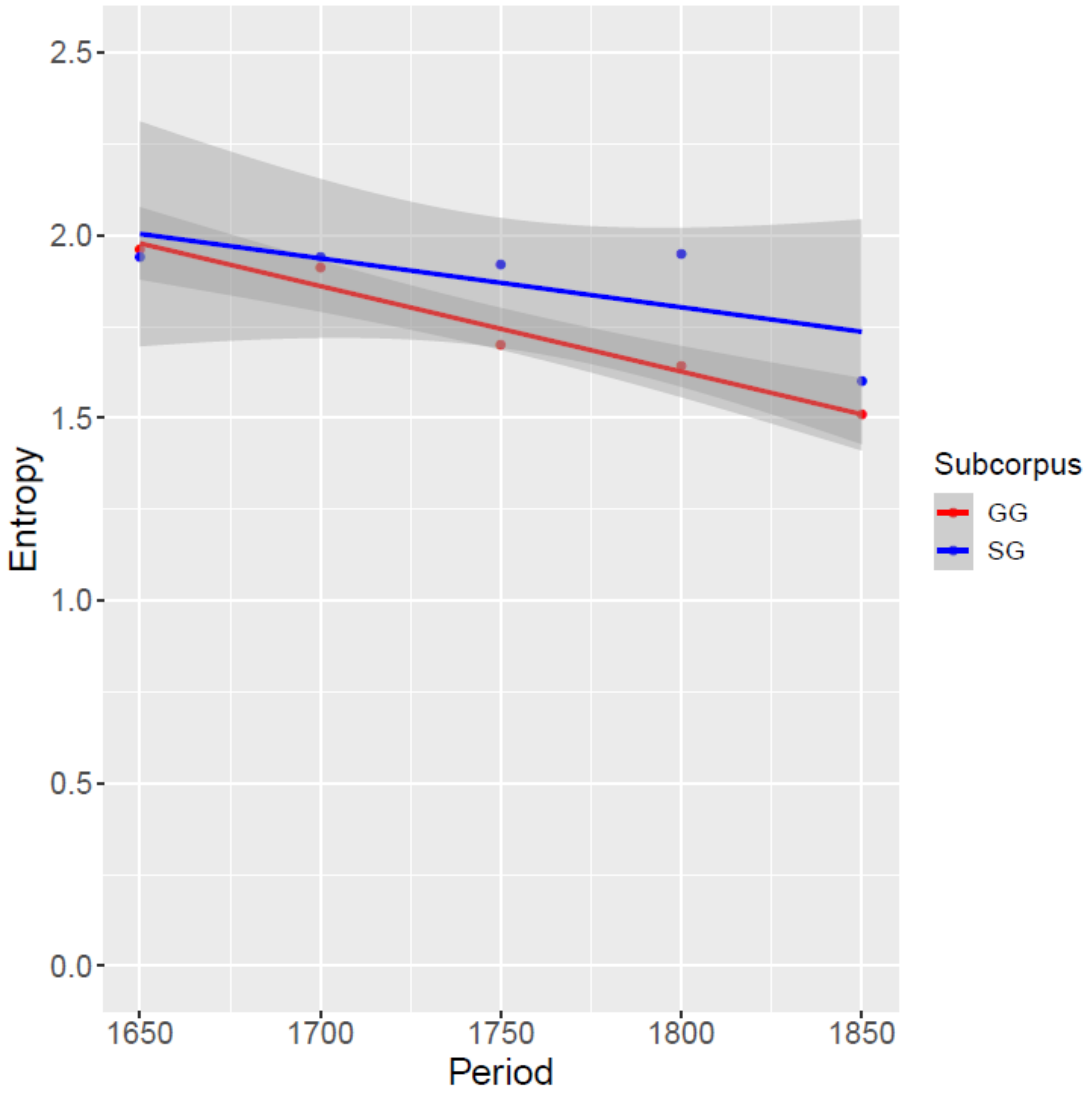

Figure 10. Entropy over the relativizer paradigm in general (GG) vs. scientific German (SG).

Figure 10 shows that entropy in GG steadily decreases after 1650, while in SG entropy is relatively stable (at approx. 1.9) until 1850 and falls after that. In 1850, entropy in SG is almost as low as in GG. During the period between 1650 and 1800, SG seems to prefer a richer choice of possible options over a monopoly of few options, whereas GG continuously develops towards a more confined set of options. Consistent with the rise in frequency of relativizers in SG until 1850, entropy, too, reflects increasing complexity regarding relativizer use and a drop thereof afterwards. For English, the results of our entropy calculations show a clear distinction between SE and GE, pointing to a clear development of a register specific preference of *which* in scientific language. In general language, however, the choice seems to be between *which* and *that*. In German, the stronger tendency of scientific texts towards diversity in relativizer choice during the period between 1650 and 1850 confirms Admoni's (1990) and Habermann's (2004) observation of expanding grammatical complexity in the scientific genre. The final drop in entropy towards 1900 reflects an eventual turn towards fewer options at the choice point of the relativizer.

5.3 Contextual predictability

We calculate surprisal for the different groups of relativizers (*which/(welch-), that/(d-)* and pronominal adverbs in order to see whether they become more or less predictable in context over time.

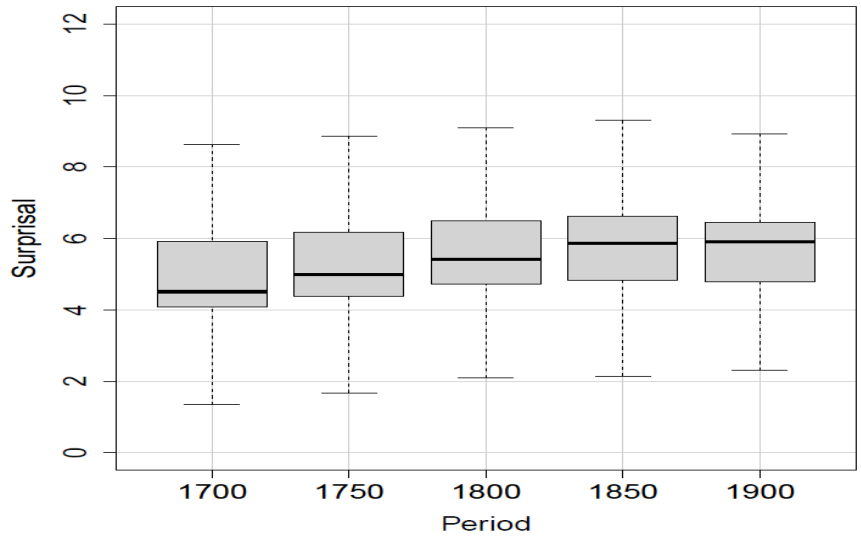

Figure 11. Distribution of surprisal values for "that" per 50 years in GE.

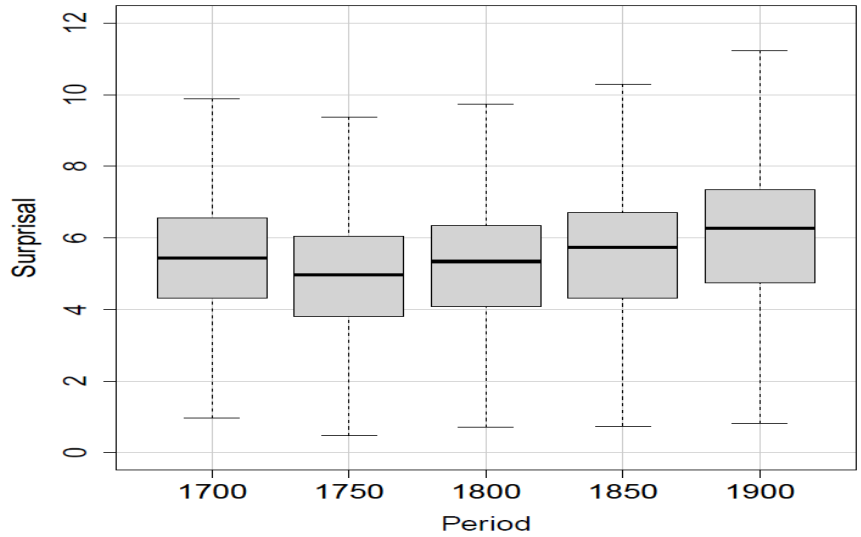

Figure 12. Distribution of surprisal values for "which" per 50 years in GE.

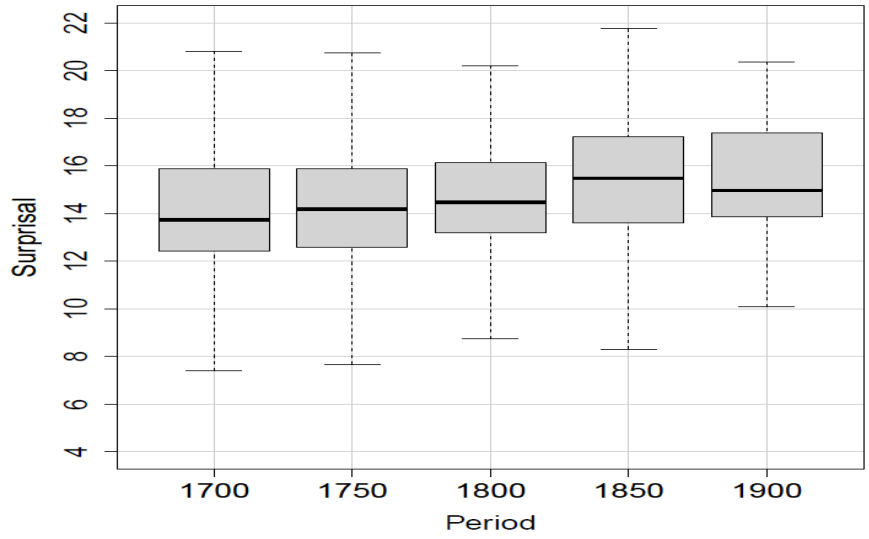

Figure 13. Distribution of surprisal values for pronominal adverbs per 50 years in GE.

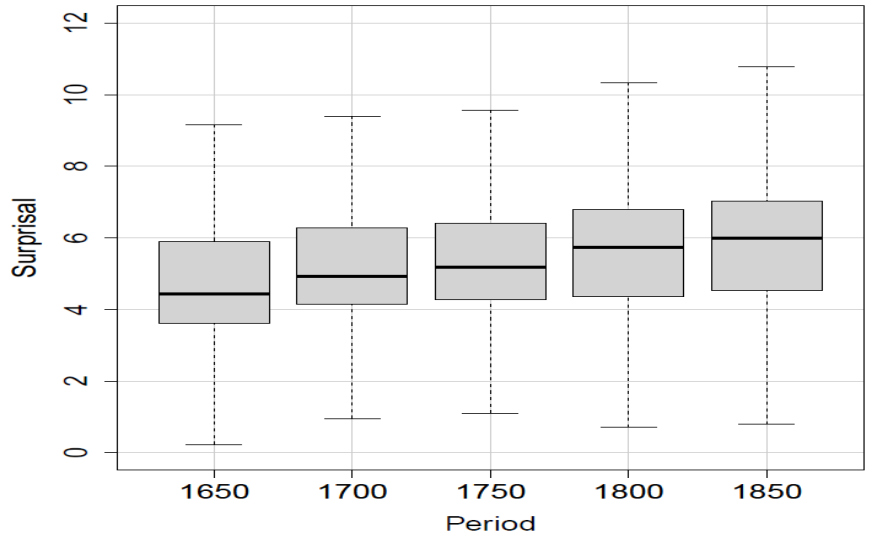

Figure 14. Distribution of surprisal values for "that" per 50 years in SE.

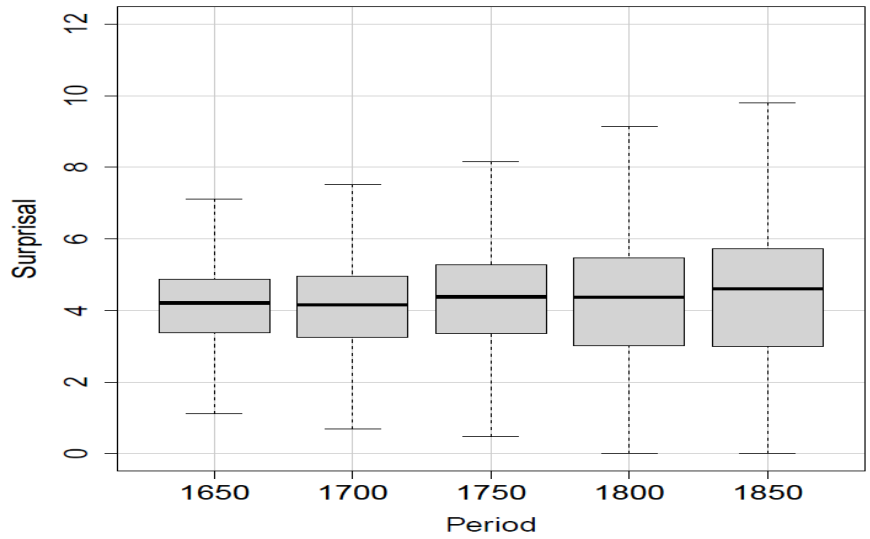

Figure 15. Distribution of surprisal values for "which" per 50 years in SE.

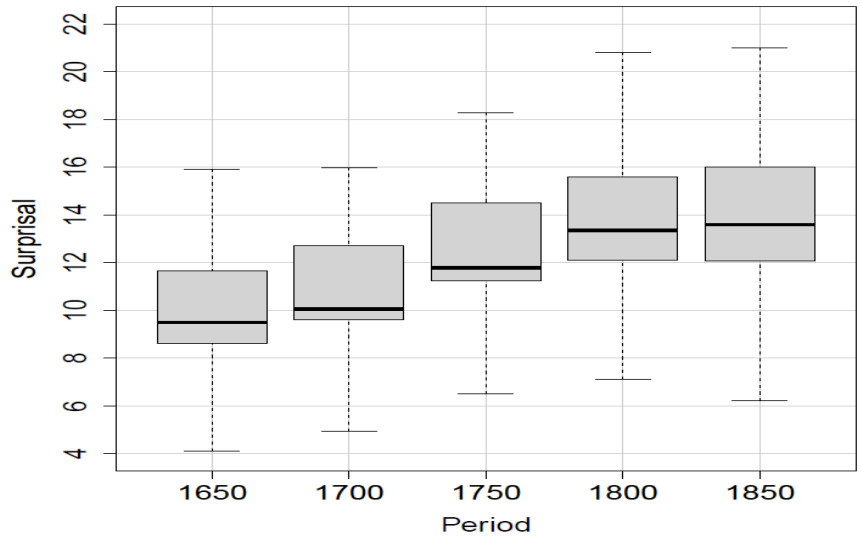

Figure 16. Distribution of surprisal values for pronominal adverbs per 50 years in SE.

The English subcorpora (figures 11−13 for GE, figures 14−16 for SE) are fairly similar for all three relativizer types. All of them become more surprising over time reflecting the decrease in use. *That* (figures 11 and 14) shows a slight increase in median surprisal values from about four to six. In GE, the maximum and minimum surprisal values (whiskers) diverge less than in SE indicating that *that* becomes more stable in terms of predictability in GE compared to *that* in SE. This is plausible since *that* is less frequently used in the latter. The surprisal values for *which* in GE (figure 12) are slightly higher than in SE (figure 15), reflecting the lower frequency in general language. In both subcorpora, the range of surprisal values increases over time. The long whiskers indicate a broader use of *which* in 1850 compared to 1650 with very frequent patterns of usage (very low surprisal values) as well as very infrequent ones (very high surprisal values). This tendency is especially evident in SE – plausibly so, since *which* over time becomes the number one relativizer in SE, while all other relativizers become less frequent. Thus, the contexts formerly covered by different relativizers are now filled by *which*. At the same time, *which* shows the most stable median surprisal values over time. This indicates that most of its preferred contexts are stable with some of them becoming particularly conventionalized and thus unsurprising. This development is in line with the theory of conventionalization in scientific language put forward by Degaetano-Ortlieb and Teich (2019) and Teich *et al.* (2021). Surprisal of pronominal adverbs in GE (figure 13) shows the constantly highest surprisal. Surprisal in SE (figure 16), instead, starts out much lower increasing continuously as pronominal adverbs and with them the possible contexts become continuously less frequent.

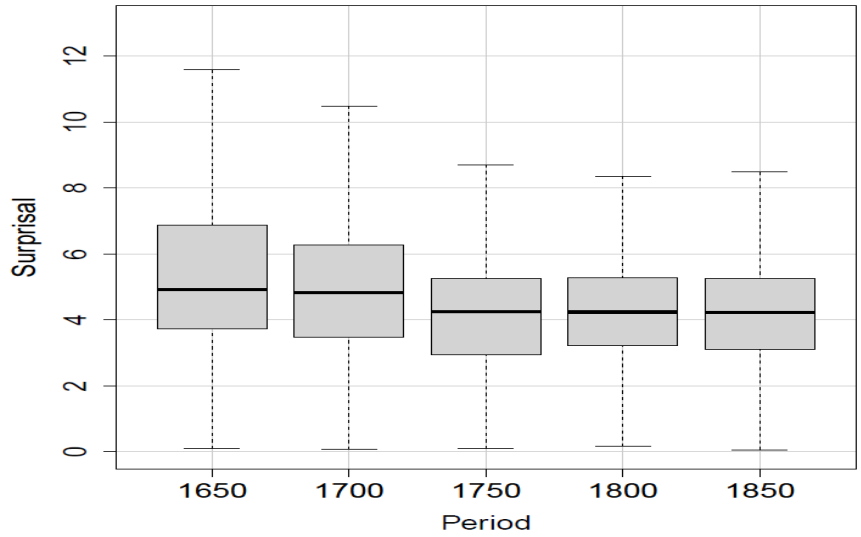

Figure 17. Distribution of surprisal values for (*d*-) per 50 years in GG.

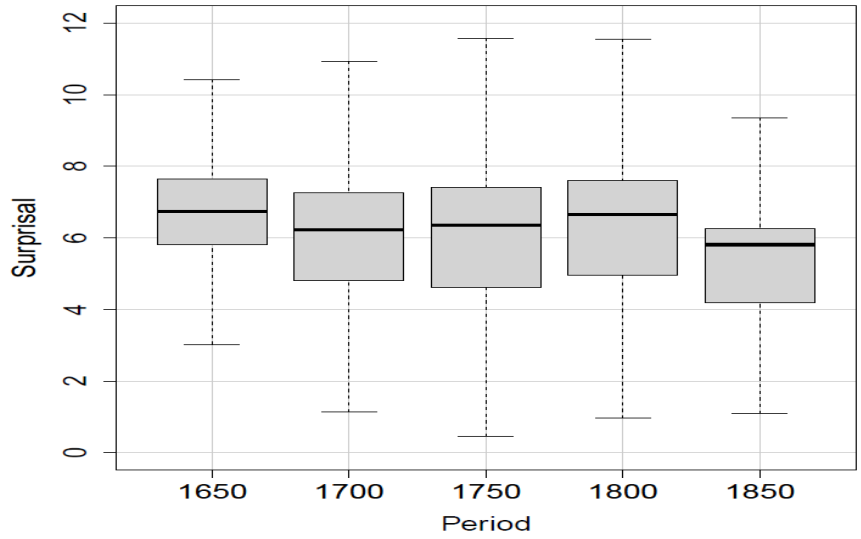

Figure 18. Distribution of surprisal values for (*welch*-) per 50 years in GG.

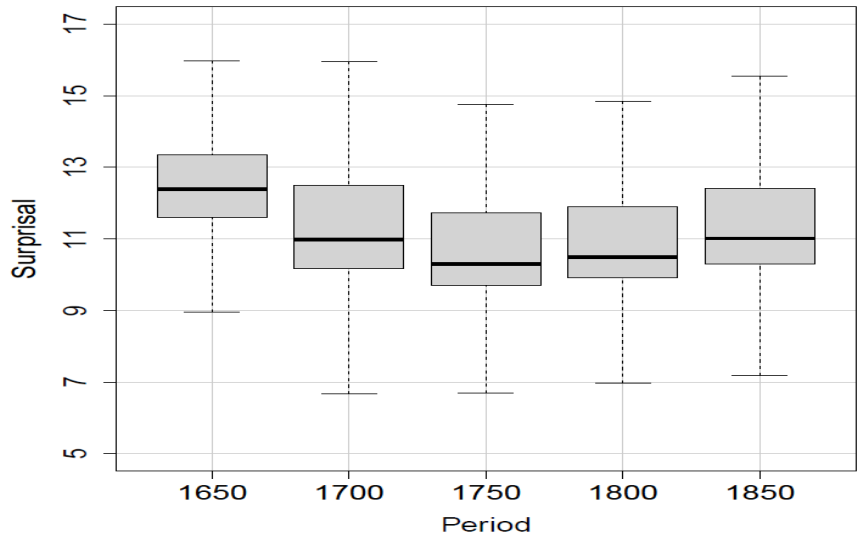

Figure 19. Distribution of surprisal values for pronominal adverbs per 50 years in GG.

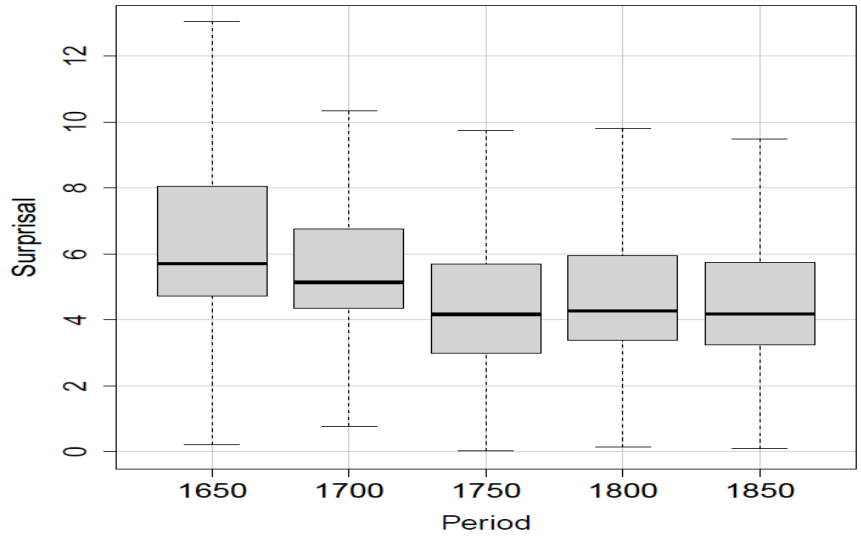

Figure 20. Distribution of surprisal values for (*d*-) per 50 years in SG.

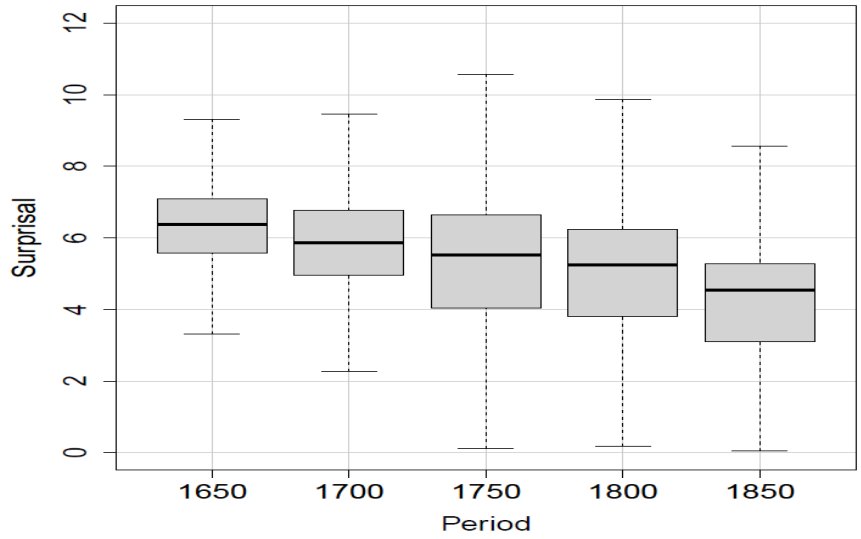

Figure 21. Distribution of surprisal values for (*welch*-) per 50 years in SG.

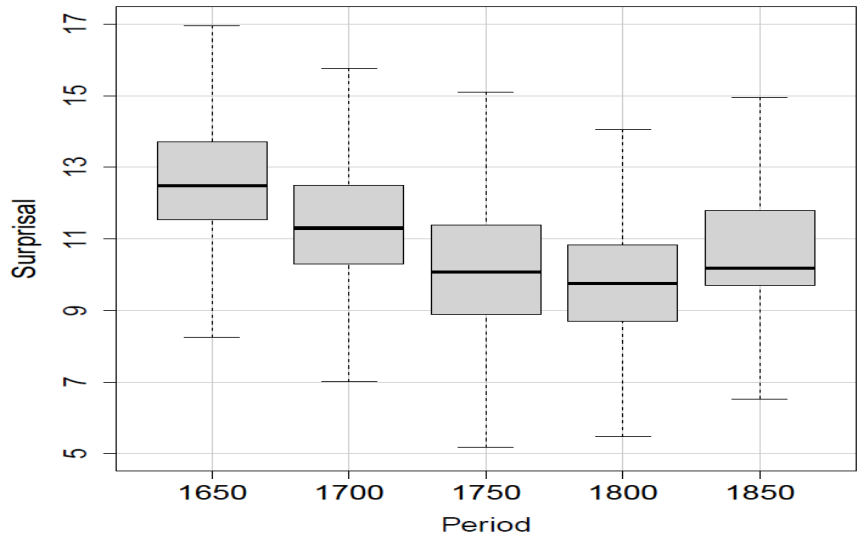

Figure 22. Distribution of surprisal values for pronominal adverbs per 50 years in SG.

In German, all relativizers become less surprising over time and the drops are steeper in scientific discourse than in general language. For the German relativizer (*d-*), contextual predictability is fairly similar between the two subcorpora (GG, figure 17 and SG, figure 20). Median surprisal values drop towards 1750 and stabilize between 1750 and 1900 at around four bits, which is the lowest surprisal value amongst all three relativizer types observed in this study. This is due to the fact that (*d-*) is the overall most frequent relativizer in both GG and SG. We can observe that the whiskers of surprisal values for (d-) in 1650 are rather broad and become narrower over time. This suggests a conventionalization of contexts of the relativizer. Looking at *(welch-)*, (figures 18 and 21), we see that the ranges of surprisal values first expand towards 1750, indicating that *(welch-)* first becomes increasingly variable in terms of contexts. This development initially seems to unfold simultaneously to its increase in frequency. After 1750, surprisal ranges become narrower, indicating a conventionalization of the contexts. In GG, the median surprisal, however, stays fairly stable over time while in SG surprisal steadily goes down. The overall frequency increase of *(welch-)* in SG obviously brings with it an increase in total contexts the relativizer occurs in. The gradual decrease in average surprisal, however, reflects that the contexts become more similar leading to easier predictability of the target word. The contextual predictability of pronominal adverbs shows a similar decrease over time, dropping lower in SG (figure 22) than in GG (figure 19), indicating conventionalization of contexts in SG. Comparing trends in German and English, we find that English relativizers overall become more surprising while in German they become less so. This is primarily due to the overall development of RCs becoming less frequent in English than in German. Only *which* in SE is stable in surprisal and seems to occur in highly predictable contexts. In the next section, we perform qualitative analyses of the syntagmatic environments of relativizers to gain further insights on the contexts RCs tend to occur in.

5.4 Syntagmatic context of relativizers

In the following qualitative analysis, we concentrate on the relativizer *which* in English and *(welch-)* in German, since these have shown to be the most distinctive ones for scientific discourse.

5.4.1 *Grammatical contexts*

To find out what contexts the relativizers occur in and whether these change over time, we first extract all part-of-speech trigrams preceding *which*[2] and *(welch-)*[3] and plot the three most frequent trigrams in each time period. Since the most frequent three in one period may overlap with trigrams from other periods, the total number of trigrams displayed varies. A low total number of trigrams in each figure indicates a lower variation between periods, while a higher number points to stronger variation. Our hypothesis here is that contexts in scientific language become more conventionalized, i.e., the frequencies of a specific pattern surpass the frequencies of other patterns.

[2] Penn Tagset: DT = determiner, NN = noun, IN = preposition, JJ = adjective

[3] STTS: ART = article, NN = noun, APPR = preposition, PT = punctuation (In earlier stages, German punctuation was not standardized yet, thus for this study all punctuation marks (<,>, <;>, </>, etc.) were normalized to 'PT' for better comparability.)

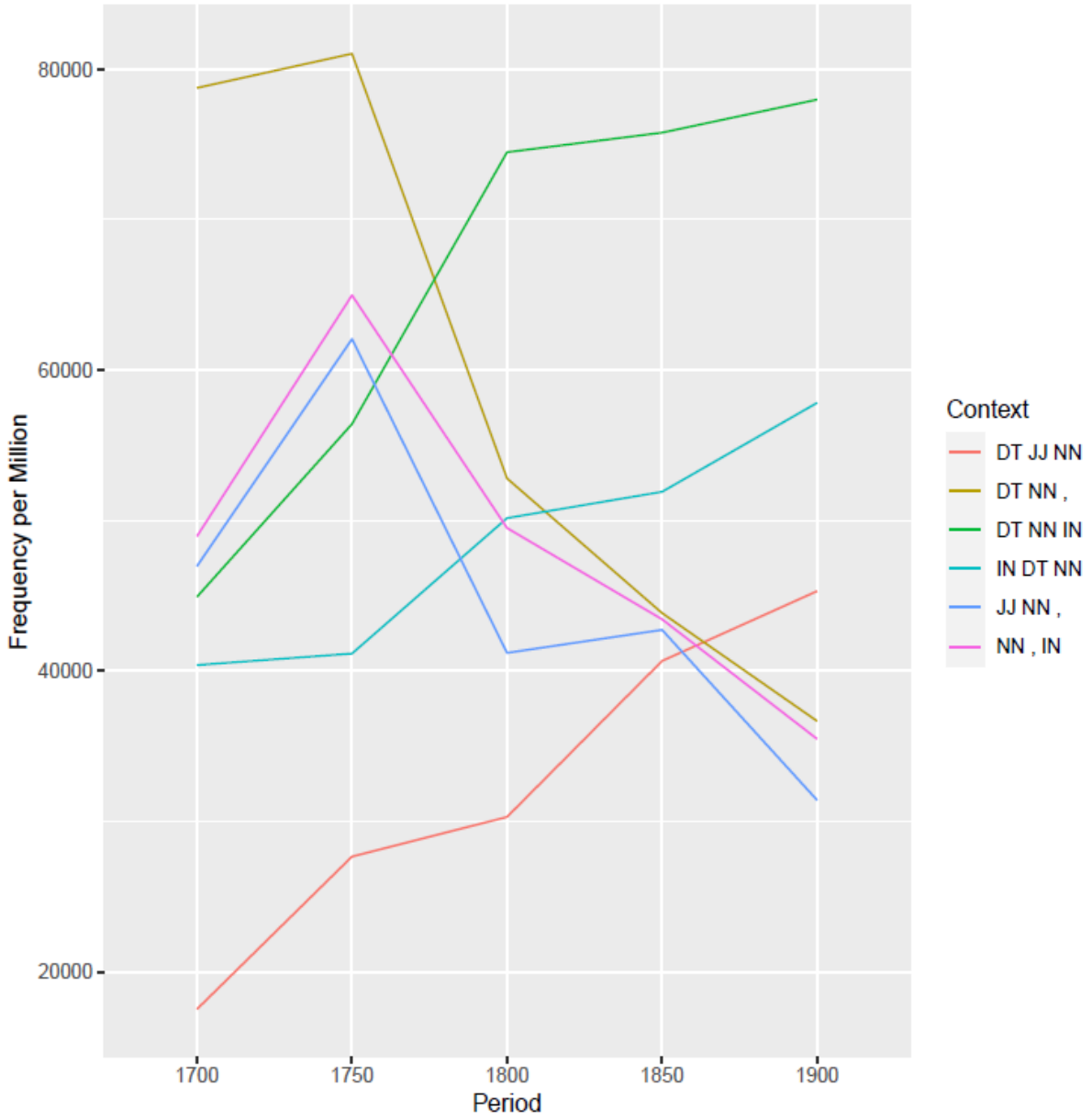

Figure 23. Three most frequent part-of-speech trigrams preceding “which” in GE per 50 years.

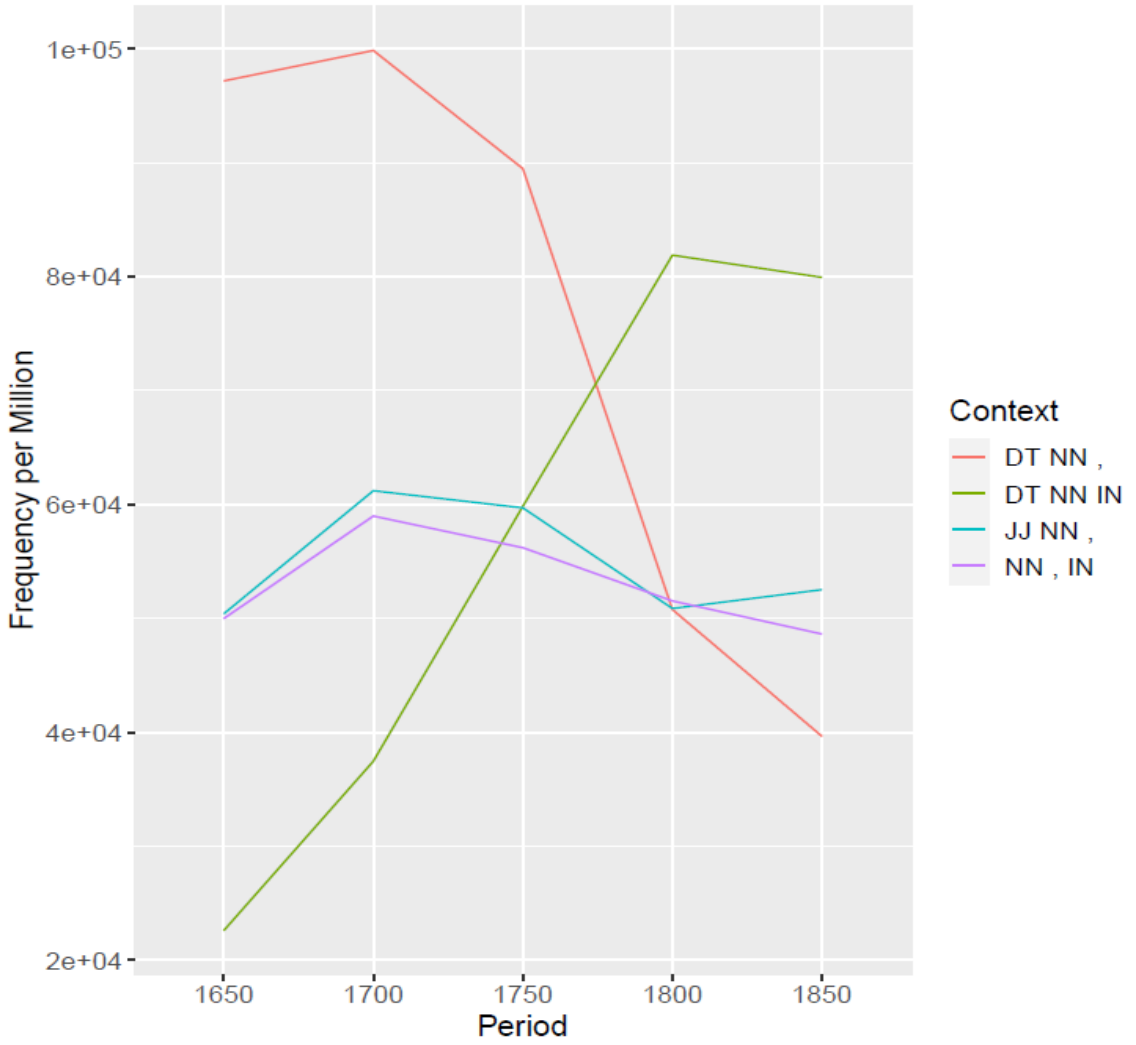

Figure 24. Three most frequent part-of-speech trigrams preceding "which" in SE per 50 years.

Comparing the trigram contexts in GE (figure 23) vs. SE contexts (figure 24), we can report a higher variation between most frequent contexts preceding relativizers in GE, as well as a more diverse set of increasing frequent contexts in GE than in SE. In the scientific corpus, all contexts decrease except for one clearly preferred pattern <*determiner noun preposition*> (DT NN IN), representing RCs introduced by a stranded preposition. Altogether, this corroborates our assumption that, in scientific discourse, contexts of RCs become increasingly conventionalized, again in line with Degaetano-Ortlieb and Teich (2019) and Teich *et al.* (2021). In German (figure 25 for GG; figure 26 for SG), the contexts preceding *(welch-)* are less diverse than in English, showing three clearly preferred contexts in both subcorpora, <adjective noun punctuation mark> (ADJA NN PT) representing shorter nominal phrases, <article noun punctuation mark> (ART NN PT) representing longer nominal phrases and <noun punctuation mark preposition> (NN PT APPR). The fact that these three patterns are continuously amongst the three most frequent POS contexts in both subcorpora alike points to a relatively rigid grammatical environment of RCs in German compared to English. Also, in both GG and SG the trajectories of the trigrams' normalized frequencies are strikingly similar until 1800, all of them increasing.

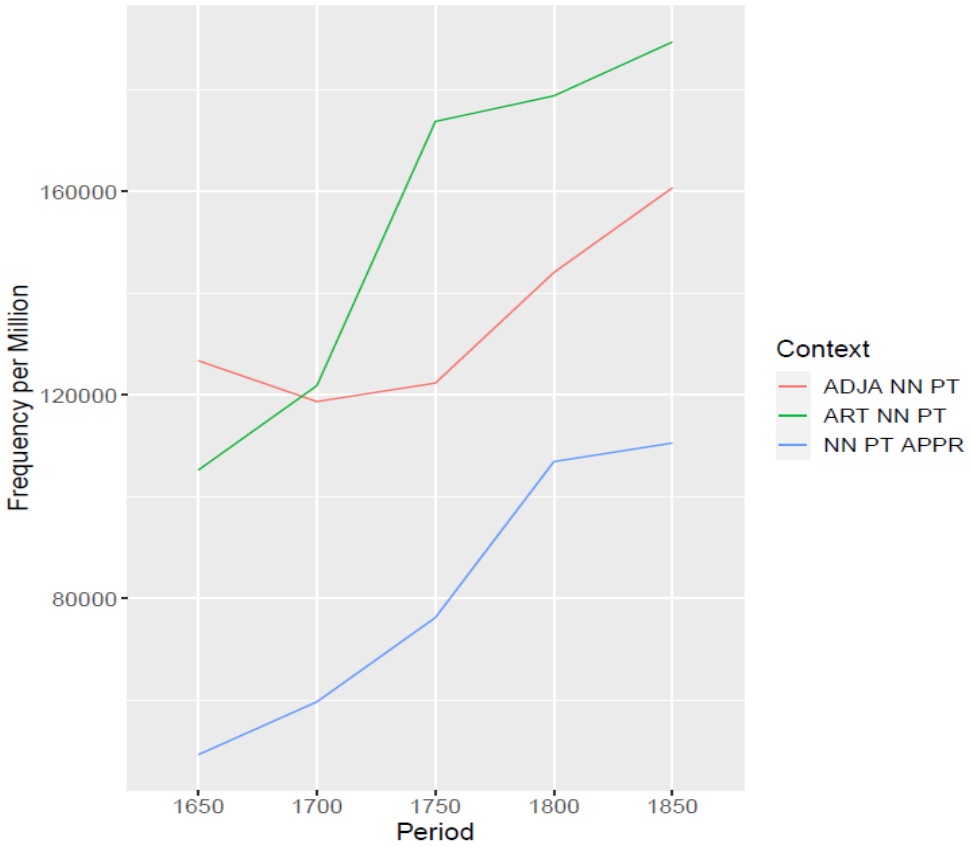

Figure 25. Three most frequent part-of-speech trigrams preceding (*welch*-) in GG per 50 years.

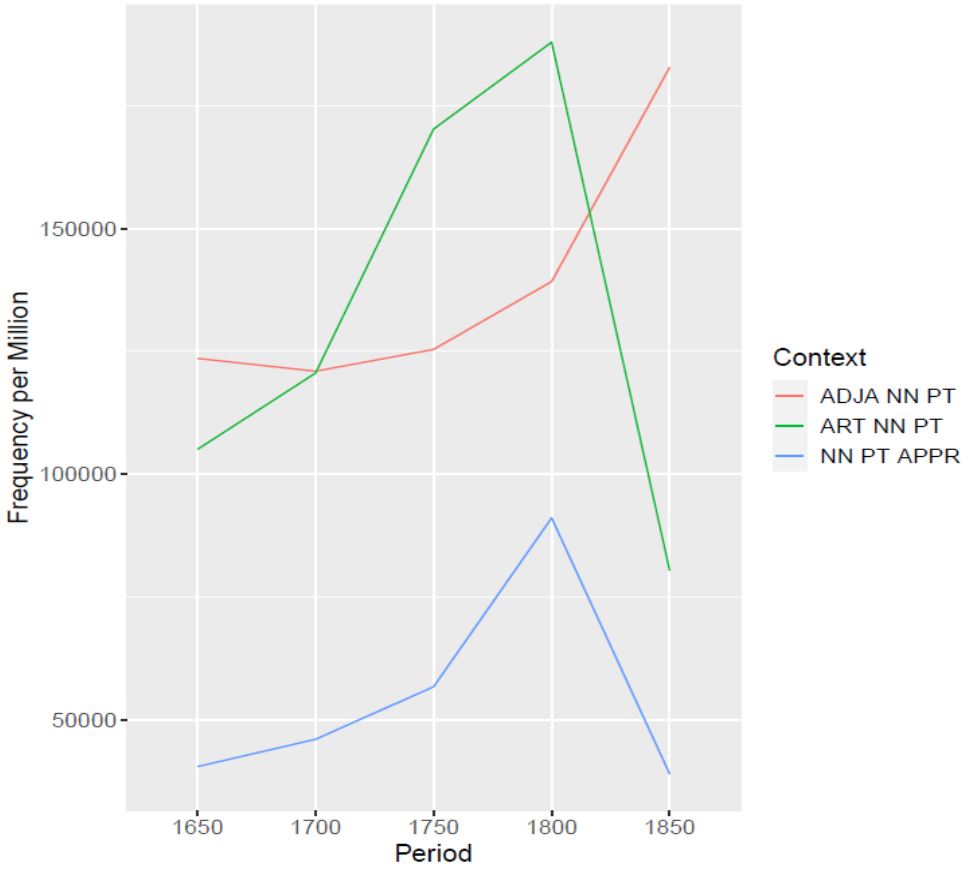

Figure 26. Three most frequent part-of-speech trigrams preceding (welch-) in SG per 50 years.

Between 1850 and 1900, however, in SG only the complex noun phrase pattern <adjective noun punctuation> (ADJA NN PT) increases, while the other two decrease. This again points to a strong differentiation in contextual use of *(welch-)*, similar to the observed trend of *which* in SE.

5.4.2 *Lexical contexts*

We further analyze the most frequent lexical trigrams preceding relativizers in both languages and subcorpora. Tables 2−5 show the three most frequent lexical trigrams per period and specific types they can be grouped in. The types occurring most often in a subcorpus are in boldface.

In English, the top three lexical trigrams preceding *which* show two clear tendencies. In GE (table 2), most top three trigrams (9/15) describe manner expressions (*the manner in*; *the way in*), forming complex conjunctions. In SE (table 4), the trigrams appearing most often are expressions of quantification (8/15) (*, out of*; *, some of*; *, one of*). In 1750 and 1850, however, the top most frequent trigram preceding *which* is the complex conjunction expressing manner, (*the manner in*), matching the top most frequent POS trigram (DT NN APPR). Interestingly, both general and scientific texts show a preference for manner expressions followed by *which*, while the exact lexical form differs between subcorpora. In 1850 GE, (*the way in*) and (*the manner in*) still compete, giving way to the first option from 1900 onwards. SE, however, adopts (*the manner in*).

Table 2. Lexical 3gram context of "which" in CLMET.

period	Freq pM	Freq raw	3-gram	type
1700	2648.87	52	, and of	partitive
	1782.89	35	the manner in	**manner**
	1579.14	31	, and to	-
1750	827.31	106	the manner in	**manner**
	556.85	82	in consequence of	causal
	493.21	78	, one of	quantification
1800	836.32	178	the manner in	**manner**
	562.91	81	, and in	-
	498.58	79	the way in	**manner**
1850	1074.60	137	the way in	**manner**
	723.29	99	the manner in	**manner**
	640.63	63	, all of	quantification
1900	3254.27	54	the way in	**manner**
	2190.37	32	the sense in	**manner**
	1940.05	19	in a way	**manner**

Table 3. Lexical 3gram context of "which" in RSC.

period	Freq pM	Freq raw	3-gram	type
1650	2882.78	62	, out of	**quantification**
	1999.35	43	the doing of	-
	1673.87	36	of it ,	-
1700	2686.05	75	, some of	**quantification**
	2471.17	69	of it ,	-
	2184.66	61	, one of	**quantification**
1750	2359.34	118	the manner in	manner
	1739.51	87	, one of	**quantification**
	1719.52	86	, some of	**quantification**
1800	4168.46	320	the manner in	manner
	1875.81	144	, one of	**quantification**
	1706.46	131	the mode in	manner
1850	2884.18	228	the manner in	manner
	1922.79	152	, each of	**quantification**
	1846.89	146	, one of	**quantification**

In German, the majority of the top three lexical trigrams preceding *(welch-)* for both registers (tables 4−5) are pronominal antecedents, i.e. (*von denen,*). The lexical trigrams do not match the most frequent POS trigrams representing noun phrases, which suggests that, in German, RCs do not tend to occur in conventionalized contexts. Apart from pronominal antecedents, we find prepositional clauses (*Zeit*, *in*; as in example (4a) and (4b) below), as well as semantically empty lexical bundles of grammatical words (*ist es,*) or (*zu sein,*). Taking a closer look at the full contexts of those verbal fragments, i.e. (*ist es,*) in the 19th century German RCs, we find the antecedents to be topicalized noun phrases in cleft constructions, as illustrated in example (5).

Table 4. Lexical 3gram context of (*welch*-) in GE.

period	Freq pM	Freq raw	3-gram	type
1650	2453.81	51	die jenige /	pronoun
	1395.30	29	Wurzel / aus	preposition
	1347.19	28	den jenigen /	pronoun
1700	4684.03	173	, Fluß,	NP
	1868.20	69	von denen,	pronoun
	1272.54	47	in England,	PP
1750	2807.97	62	zu machen,	*to*-infinitive
	1856.88	41	. Diejenigen,	pronoun
	1585.14	35	als die,	pronoun
1800	1625.54	33	als die,	pronoun
	1428.50	29	Zeit, in	preposition
	1083.69	22	ist, in	preposition
1850	1661.13	42	Zeit, in	preposition
	1067.87	27	ist es,	cleft
	909.67	23	Tage, an	preposition

Table 5. Lexical 3gram context of (*welch*-) in SE.

period	Freq pM	Freq raw	3-gram	type
1650	3528.29	52	die jenige /	pronoun
	1899.85	28	die jenigen /	pronoun
	1832.00	27	der Linie /	NP
1700	1413.27	38	Tochter, mit	preposition
	1264.50	34	daß diejenigen,	pronoun
	1115.74	30	zu sehen,	*to*-infinitive
1750	2644.88	157	. Diejenigen,	pronoun
	1212.94	72	zu sein,	*to*-infinitive
	1179.25	70	als die,	pronoun
1800	2967.13	157	. Diejenigen,	pronoun
	1360.72	72	zu sein,	*to*-infinitive
	1322.93	70	als die,	pronoun
1850	1567.25	194	ist es,	cleft
	1147.16	142	Zeit, in	preposition
	1009.82	125	sind es,	cleft

The trigram (*zu sein,*), illustrated in example (6), derives from *scheinen* + *zu*-infinitive constructions (Engl. *seem* + *to*-infinitive).

(4)

a) Nach dem vorigen ist die *Zeit, in welcher* das ganze Gefäß ausfließt, T = [formula]. (gerstner_mechanik02_1832, Scientific German)

b) Endlich nahte die *Zeit, in welcher* man in den Sternenhof gehen sollte. (stifter_nachsommer02_1857, General German)

(5) Gerade diese Zugehörigkeit zu einer und derselben Gruppe von Vorstellungen ist es, *welche* hier die Annahme von Ähnlichkeitsassoziationen rechtfertigt. (kraepelin_arzneimittel_1892)

(6) Es scheint mir ferner eine berechtigte Auffassung zu sein, *welche* Darwin in einem trefflichen Beispiele ausspricht […] (roux_kampf_1881)

Overall, *(welch-)* shares similar lexical contexts in both subcorpora with a slight preference for cleft- constructions in SG and a slight preference for prepositional phrases in GG. The fact that lexical trigrams do not map the POS trigrams shows that, in German, RCs are not introduced by lexicalized multi-word units. Instead, they often occur in frequent syntactic constructions such as *to*-infinitives and topicalization in cleft-constructions. The decreasing surprisal values of German relativizers (which are calculated on lexical patterns) are most likely attributable to the increasingly conventionalized use of comma introducing relative clauses.

6. Summary and discussion

In this paper, we have conducted a comparative study on German and English relativizers as indicators of grammatical complexity. Specifically, we pursued the hypothesis that scientific texts become grammatically less complex compared to texts from general language. We tested for complexity in terms of syntactic intricacy, paradigmatic richness and contextual predictability of relativizers.

Our hypothesis that RCs become less frequent in scientific language is confirmed. However, this development is not exclusive to scientific language but rather concerns all subcorpora. While the decrease in English follows a linear trend, in German RC frequency first increases immensely until the second half of the 18th century and only decreases afterwards, perfectly in line with descriptions by Möslein (1974) and Beneš (1981).

In terms of embeddedness, we found that in English indeed, the average number of RC embeddings per sentence decreases proportionally to the overall number of relativizers in a 50 years' period. The trend confirms Halliday and Martin's (1993) claims about a reduction in syntactic intricacy in scientific English. In German, we found that embeddedness is overall stronger in general German than in scientific discourse. Embeddedness in German is highest before RC frequencies reach their climax, indicating a trend towards a more balanced subordination over time. Overall, a comparison of mere frequencies did not show a register specific trend for lower syntactic intricacy in terms of RC use in the observed time periods, but rather a general linguistic development during the time between 1650 and 1850.

In terms of paradigmatic richness, we found that scientific English developed from a richly populated paradigm of manifold relativizers (especially pronominal adverbs) in 1650 towards a clear preference for one single relativizer, *which*, in 1850. General English, in contrast, remained stable, showing broadly the same distributions of relativizers across all time periods. Calculating entropy, we found that scientific English shifted towards an extremely low uncertainty about an upcoming relativizer, which confirms Degaetano-Ortlieb and Teich's (2019) theory of conventionalization. For German, we found an inverse development. General German develops towards an increasingly confined choice, prioritizing (*d-*) and continuously decreasing in entropy, while scientific German shows a much broader choice of relativizers and consistently high entropy until 1850. In the second half of the 19th century, we observe a drop in use of pronominal adverbs ultimately leading to a decrease in entropy between 1850 and 1900. The findings also show that scientific German over a long stretch of time prefers a rich paradigmatic choice for sophisticated expression, while in English scientific texts a smaller set of choices is preferred. Again, between 1850 and 1900 trends in the two scientific subcorpora align.

Regarding contextual predictability, for English we found overall increasing surprisal for all relativizers in both subcorpora, while in German surprisal values go down. This general result reflects the development of relativizer frequency in the two languages. Obviously, when relativizers overall become less frequent, like in English, they become less predictable. At the same time, *which* becomes least surprising in scientific texts, confirming that during register formation certain words become conventionalized in specific contexts. In German, we see an inverse development. All relativizers become more predictable due to their increasing frequency over time. However, we observed steeper drops in surprisal in scientific texts, especially for the relativizer *(welch-)*. Surprisal values of *(welch-)* also show a smaller range indicating increasingly conventionalized contexts of the relativizer most strongly associated with scientific discourse.

Our qualitative comparison of grammatical and lexical contexts of the relativizers *which* and *(welch-)* showed that in English the most frequent grammatical and lexical contexts of *which* overlap and represent highly lexicalized multi-word units (i.e. (DT NN IN) expressing manner and quantification (*the manner in which, one of which*). In German, the most frequent grammatical contexts do not match with most frequent lexical contexts, indicating that grammatical contexts in German do not become lexicalized over time. The most frequent lexical contexts rather reflect common grammatical constructions of the time, such as topicalized cleft-constructions (*Die Frau war es, welche den Mann schlug.*) and *to*-infinitives in epistemic phrases (*scheint eine Frau zu sein, welche…*) typical for scientific discourse.

Overall, in the scientific subcorpora, we found largely inverse developments in English (becoming less complex) and German (becoming more complex) until the first half of the 19th century and an alignment towards lower complexity in the second half. The results are in line with related work (Möslein, 1974; Beneš, 1981; Admoni, 1990) and our hypothesis that grammatical complexity in German should decrease much later than in English. The delayed shift towards lower complexity in German scientific language may be due to several factors,

such as the longstanding Latin influence on German linguistic style (cf. Habermann, 2001), as well as a much later institutional implementation of German scientific discourse and ultimately a language specific preference for explicit style as compared to English (cf. House, 2006).

Acknowledgements

I am grateful to the anonymous reviewers for their detailed comments and suggestions for improving the paper.

References

Aarts, B., López-Couso, M.J. and Méndez-Naya, B. 2012. Late modern English syntax. In *Historical Linguistics of English*, A. Bergs and L.J. Brinton (eds), 869–887. Berlin/Boston: Mouton de Gruyter.

Admoni, W. 1990. *Historische Syntax des Deutschen*. Tübingen: Niemeyer.

Ágel, V. 2000. Syntax des Neuhochdeutschen bis zur Mitte des 20. Jahrhunderts. In *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung*, W. Besch, A. Betten, O. Reichmann and S. Sonderegger (eds), 1855–1903. Berlin: De Gruyter.

Ball, C. 1996. A Diachronic Study of Relative Markers in Spoken and Written English. *Language Variation and Change* 8 (2): 227−258.

Beneš, E. 1981. Die formale Struktur der wissenschaftlichen Fachsprachen aus syntaktischer Hinsicht. In *Wissenschaftssprache*, T. Bungarten (ed.), 185−212. München: Fink.

Betten, A. 2016. *Grundzüge der Prosasyntax*. Berlin/Boston: Max Niemeyer Verlag.

Biber, D. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. 1993. The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings. *Computers and the Humanities* 26 (5-6): 331−345.

Biber, D. 2006. *University Language: A Corpus-based Study of Spoken and Written Registers.* Amsterdam: John Benjamins Publishing.

Biber, D. 2012. Register as a Predictor of Linguistic Variation. *Corpus Linguistics and Linguistic Theory* 8 (1): 9−37.

Biber, D. and Clark, V. 2002. Historical Shifts in Modification Patterns with Complex Noun Phrase Structures. In *English Historical Syntax and Morphology. Selected Papers from 11 ICEHL, Santiago de Compostela 2002*, T. Fanego, J. Pérez-Guerra and M.J. López-Couso (eds). 43−66.

Biber, D. and Conrad, S. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Biber, D. and Finegan, E. 1997. Diachronic Relations among Speech-based and Written Registers in English. In *To explain the present: Studies in the changing English language in honour of Matti Rissanen*, T. Nevalainen and L. Kahlas-Tarkka (eds), 253−275. Helsinki: Modern Language Society.

Biber, D. and Gray, B. 2011. Grammatical Change in the Noun Phrase: The Influence of Written Language Use. *English Language and Linguistics* 15 (2): 223−250.

Biber, D. and Gray, B. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing.* Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Brooks, T. 2006. *Untersuchungen zur Syntax in oberdeutschen Drucken des 16.–18. Jahrhunderts*. Frankfurt a.M.: Lang.

Crocker, M.W., Demberg, V. and Teich, E. 2015. Information Density and Linguistic Encoding (IDeaL). *KI - Künstliche Intelligenz* 30 (1): 77−81.

Dal, I. 2014. *Kurze deutsche Syntax auf historischer Grundlage*. Berlin: De Gruyter.

Degaetano-Ortlieb, S., Kermes, H, Khamis, A. and Teich, E. 2016. An Information-Theoretic Approach to Modeling Diachronic Change in Scientific English. In *Selected Papers from Varieng − From*

Data to Evidence, *Language and Computers*, C. Suhr, T. Nevalainen and I. Taavitsainen (eds), 258−281. Leiden: Brill.
Degaetano-Ortlieb, S. and Teich, E. 2016. Information-based Modeling of Diachronic Linguistic Change: From Typicality to Productivity. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 165−173.
Degaetano-Ortlieb, S. and Teich, E. 2019. Toward an Optimal Code for Communication: The Case of Scientific English. *Corpus Linguistics and Linguistic Theory*. Available at https://www.degruyter.com/document/doi/10.1515/cllt-2018-0088/html [Last accessed 2 June 2021].
Diller, H., De Smet, H. and Tyrkkö, J. 2011. A European Database of Descriptors of English Electronic Texts. *The European English Messenger* 19: 21−35.
Ebert, R.P. 1986. *Historische Syntax des Deutschen II: 1300–1750.* Bern: Lang.
Fleischer, J. 2004. A Typology of Relative Clauses in German Dialects. In *Trends in Linguistics. Dialectology Meets Typology. Dialect Grammar from a Cross-linguistic Perspective*, B. Kortmann (ed.), 211–243. Berlin/New York: De Gruyter Mouton.
Geyken, A., Boenig, M., Haaf, S., Jurish, B., Thomas, C. and Wiegand, F. 2018. Das Deutsche Textarchiv als Forschungsplattform historische Daten in CLARIN. In *Digitale Infrastrukturen die germanistische Forschung (= Germanistische Sprachwissenschaft um 2020, Bd. 6*), H. Lobin, R. Schneider and A. Witt (eds), 219–248. Berlin/Boston: De Gruyter.
Görlach M. 2004. *Text Types and the History of English*. Berlin/New York: Mouton de Gruyter.
Guy, G. and Bayley, R. 1995. On the Choice of Relative Pronouns in English. *American Speech* 70 (2): 148–162.
Habermann, M. 2011. *Deutsche Fachtexte der frühen Neuzeit*. Berlin/Boston: De Gruyter.
Halliday, M.A.K. and R. Hasan. 1985. *Language, Context, and Text: Aspects of Language in a Social-semiotic Perspective.* Oxford: Oxford University Press.
Halliday, M.A.K. 1988. On the Language of Physical Science. In *Registers of Written English: Situational Factors and Linguistic Features*, M. Ghadessy (ed.), 162−177. London: Pinter.
Halliday, M.A.K and Martin, J.R. 1993. *Writing Science: Literacy and Discursive Power*. London: Falmer Press.
Hinrichs, L., Szmrecsanyi, B. and Bohmann, A. 2015. Which-hunting and the Standard English RC. *Language* 91 (4): 806–836.
House, J. 2006. Communicative Styles in English and German. *European Journal of English Studies* 10 (3): 249−267.
Hundt, M., Denison, D. and Schneider, G. 2012. Relative Complexity in Scientific Discourse. *English Language and Linguistics* 16 (2): 209–240.
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*, Barcelona, Spain.
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*. Portorož, Slovenia.
Krielke, M.-P., Fischer, S., Degaetano-Ortlieb, S. and Teich, E. 2019. System and Use of *wh*-relativizers in 200 years of English Scientific Writing. In *Proceedings of the 10th International Corpus Linguistics Conference 2019.* Cardiff, Wales, UK.
Leech, G., Hundt, M., Mair, C. and Smith, N. 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.
Lehmann, H. 2001. Zero Subject Relative Constructions in American and British English. *Language and Computers* 36 (1): 163–177.
Levey, S. 2006. Visiting London Relatives. *English World-Wide* 27 (1): 45–70.
Levy, R. 2008. Expectation-based Syntactic Comprehension. *Cognition* 106 (3): 1126−1177.
Mair, C. 2006. *Twentieth-century English: History, Variation and Standardization*. Cambridge: Cambridge University Press.
Mellinkoff, D. 2004. *The Language of the Law*. Eugene: Resource Publications.
Milin, P., Kuperman, V., Kostic, A. and Baayen, H.. 2009. Paradigms Bit by Bit: An Information Theoretic Approach to the Processing of Paradigmatic Structure in Inflection and Derivation. In

Analogy in Grammar: Form and Acquisition, J.P. Blevins and J. Blevins (eds), 214−252. Oxford: Oxford University Press.

Möslein, K. 1974. Einige Entwicklungstendenzen in der Syntax der wissenschaftlich technischen Literatur seit dem Ende des 18. Jahrhunderts. *Beiträge zur Geschichte der deutschen Sprache und Literatur* 94: 156−198.

Nevalainen, T. 2012. Reconstructing Syntactic Continuity and Change in Early Modern English Regional Dialects: The Case of *who*. In *Analyzing Older English*, D. Denison, R. Otero, C. McCully and E. Moore (eds), 159–184. Cambridge: Cambridge University Press.

Nevalainen, T. and Raumolin-Brunberg, H. 2002. The Rise of Relative *who* in Early Modern English. In *Relativisation on the North Sea Littoral*, P. Poussa (ed.), 109–121. Munich: Lincom Europa.

Nevalainen, T. and Raumolin-Brunberg, H. 2012. Its Strength and the Beauty of it: The Standardization of the Third Person Neuter Possessive in Early Modern English. In *Towards a Standard English*, D. Stein and I. Tieken-Boon van Ostade (eds), 171−216. Berlin/Boston: De Gruyter Mouton.

Österman, A. 1997. *There*-compounds in the History of English. *Topics in English Linguistics* 24: 191−276.

Pickl, S. 2020. Factors of Selection, Standard Universals, and the Standardisation of German Relativisers. *Lang Policy 19:* 235−258.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Reichmann, O. and Wegera, K.-P. 1993. *Frühneuhochdeutsche Grammatik*. Tübingen: Niemeyer.

Romaine, S. 1980. The RC Marker in Scots English: Diffusion, Complexity, and Style as Dimensions of Syntactic Change. *Language in Society* 9 (2): 221–247.

Romaine, S. 1982. *Sociolinguistic Variation in Speech Communities*. London: Edward Arnold.

Rubino, R., Degaetano-Ortlieb, S., Teich, E., and van Genabith, J. 2016. Modeling Diachronic Change in Scientific Writing with Information Density. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*: 750−761.

Santorini, B. 1990. *Part-of-speech Tagging Guidelines for the Penn Treebank Project* (3rd revision). Technical Report MS-CIS-90-47, University of Pennsylvania, Department of Computer and Information Science.

Shannon, C.E. 1949. *The Mathematical Theory of Communication*. Urbana/Chicago: University of Illinois Press, 1983 edition.

Tagliamonte, S. 2002. Variation and Change in the British Relative Marker. In *Relativisation on the North Sea Littoral*, P. Poussa (ed.), 147–165. Munich: Lincom Europa.

Tagliamonte, S., Smith, J. and Lawrence, H. 2005. No Taming the Vernacular! Insights from the Relatives in Northern Britain. *Language Variation and Change* 17 (1): 75–112.

Teich, E., Degaetano-Ortlieb, S., Fankhauser, P., Kermes, H. and Lapshinova-Koltunski, E. 2016. The Linguistic construal of Disciplinarity: A Data Mining Approach Using Register Features. *Journal of the Association for Information Science and Technology (JASIST)* 67 (7): 1668−1678.

Teich, E., Fankhauser, P., Degaetano-Ortlieb, S. and Bizzoni, Y. 2021. Less is More/More Diverse: On the Communicative Utility of Linguistic Conventionalization. *Frontiers in Communication* 5. https://doi.org/10.3389/fcomm.2020.620275 [Last accessed 2 June 2021].

Thielen, C., Schiller, A., Teufel, S. and Stöckert, C. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf [Last accessed 2 June 2021].

Tottie, G. and Harvie, D. 2000. It's All Relative: Relativization Strategies in Early African American Vernacular English. In *The English History of African American English*, S. Poplack (ed.), 198–230. Oxford: Blackwell.

Voigtmann, S. and Speyer, A.. 2020. *Information Density as a Factor for Syntactic Variation in Early New High German*, LE, Tübingen.

Von Polenz, P. 1999. *Deutsche Sprachgeschichte vom Spätmittelalter bis zur Gegenwart* (Vol. 3). Berlin/New York: Walter de Gruyter.

Author's address

Marie-Pauline Krielke
Department of Language Science and Technology
Saarland University
Building A2.2, room 1.02
Campus
DE-66123 Saarbrücken
Germany
mariepauline.krielke@uni-saarland.de