Optimizing Scientific Communication: The Role of Relative Clauses as Markers of Complexity in English and German Scientific Writing between 1650 and 1900

Dissertation
zur Erlangung des akademischen Grades
eines Doktors der Philosophie
der Philosophischen Fakultät
der Universität des Saarlandes

vorgelegt von
MARIE-PAULINE KRIELKE
aus SCHWELM

Saarbrücken, 18. Mai 2023

Dekanin der Fakultät P: Prof. Dr. Stefanie Haberzettl
Erstberichterstatterin: Prof. Dr. Elke Teich
Zweitberichterstatterin: PD Dr. Stefania Degaetano-Ortlieb
Tag der letzten Prüfungsleistung: 11. Juli 2023

Acknowledgements

I would like to express my heartfelt gratitude to the following individuals and institutions for their invaluable support, guidance, and encouragement throughout the journey of completing my dissertation:

I am deeply thankful to my Dissertation Advisor, Professor Elke Teich. Thank you for your unwavering support, expertise, and mentorship. Your guidance was instrumental in shaping my research and helping me navigate the challenges *of* scientific writing *about* scientific writing. Thank you for everything.

I extend my appreciation to my Dissertation Committee Members, Professor Elke Teich, PD Dr. Stefania Degaetano-Ortlieb, Professor Augustin Speyer, Dr. Robin Lemke and Professor Ingo Reich, for their valuable insights, constructive feedback, and dedication to my academic success.

My family has been my pillar of support throughout this process. First of all, thank you, Sebastian, for your unwavering support, patience and daily kindness. I thank my parents, my sister and my partner's parents for their unwavering belief in me and for their patience, understanding, and encouragement during the long hours of research and writing. I am aware that my frequent absence (physical and mental) from family life has not been easy, and I am deeply thankful for bearing with me until this very moment.

I am grateful to my colleagues Heike, Katja, Katrin, Koel, Jörg, Luigi, Stefan, Tom and Yuri, who provided moral and professional support, shared valuable resources, and offered a listening ear during both the highs and lows of my academic journey. I am deeply grateful to have (had) the privilege of working with such exceptional colleagues like you.

I extend my gratitude to the individuals who participated in my research, without whom this dissertation would not have been possible: Stefania, for your incredibly thoughtful and precise feedback and your brilliant ideas. Jörg, for helping me separate "Gute" from "Schlechte Sätze" and for your mathematical expertise, Luigi and Stefan for your generous and patient support in corpus creation and other computational tasks. Marius Mosbach, thank you for your patient advice in surprisal modelling issues. Without all of you, this thesis would not have been completed. Your contributions were invaluable.

I acknowledge the financial support provided by SFB 1102 for funding my research and enabling me to carry out this study.

I thank Saarland University and the Language Science and Technology Department for providing a conducive academic environment and access to resources necessary for my research.

I extend my heartfelt thanks to the remarkable women who inspired and guided me on this journey. Professors Stella Neumann and Professor Kerstin Kunz played pivotal roles in igniting my passion for contrastive corpus linguistics when they offered

me the opportunity to work as a student assistant. Professor Ekaterina Lapshinova-Koltunski deserves my deepest gratitude for playing a key role in guiding me towards an academic career. Katja, You remain a shining example of determination, resilience, and self-confidence. I am forever grateful to PD Dr. Stefania Degaetano-Ortlieb. Thank you for your unwavering belief in me, your invaluable professional support, and, above all, for being the wonderful person that you are. Thank you, from the bottom of my heart.

Completing this dissertation has been a challenging yet rewarding endeavour, and your support has made it possible. I am deeply thankful for the trust you placed in me and the contributions you made to this work. Thank you all for being a part of this important chapter in my academic journey.

Marie-Pauline Krielke
October 2023

Abstract

The aim of this thesis is to show that both scientific English and German have become increasingly optimized for scientific communication from 1650 to 1900 by adapting the usage of relative clauses as markers of grammatical complexity. While the lexico-grammatical changes in terms of features and their frequency distribution in scientific writing during this period are well documented, in the present work we are interested in the underlying factors driving these changes and how they affect efficient scientific communication. As the scientific register emerges and evolves, it continuously adapts to the changing communicative needs posed by extra-linguistic pressures arising from the scientific community and its achievements.

We assume that, over time, scientific language maintains communicative efficiency by balancing lexico-semantic expansion with a reduction in (lexico-)grammatical complexity on different linguistic levels. This is based on the idea that linguistic complexity affects processing difficulty and, in turn, communicative efficiency. To achieve optimization, complexity is adjusted on the level of lexico-grammar, which is related to expectation-based processing cost, and syntax, which is linked to working memory-based processing cost.

We conduct five corpus-based studies comparing English and German scientific writing to general language. The first two investigate the development of relative clauses in terms of lexico-grammar, measuring the paradigmatic richness and syntagmatic predictability of relativizers as indicators of expectation-based processing cost. The results confirm that both levels undergo a reduction in complexity over time. The other three studies focus on the syntactic complexity of relative clauses, investigating syntactic intricacy, locality, and accessibility. Results show that intricacy and locality decrease, leading to lower grammatical complexity and thus mitigating memory-based processing cost. However, accessibility is not a factor of complexity reduction over time.

Our studies reveal a register-specific diachronic complexity reduction in scientific language both in lexico-grammar and syntax. The cross-linguistic comparison shows that English is more advanced in its register-specific development while German lags behind due to a later establishment of the vernacular as a language of scientific communication.

Contents

Part I

Introduction and Background

Chapter 1

Introduction

In this thesis, we investigate the cross-lingual diachronic development of linguistic complexity in scientific writing in English and in German. We trace this development in scientific and general language corpora spanning the time between 1650 and 1900, which is a time of great advances in the scientific community. The scientific revolution had just been concluded and scientific academies were born: the *Royal Society* in Britain and the *Leopoldina* in Germany. The formation of these new institutions had a great impact, not only on giving an institutional framework to academic efforts and a concentration of forces, but also on the linguistic standardization of the vernacular languages newly becoming the languages of scientific communication. Until that time, Latin had been the lingua franca of science in the European scientific community; however, humanist culture and technical necessity led to opening up science to the public. When the vernacular language was introduced for official use in science, there was also a growing interest in standardizing it. Thus, dictionaries, as well as reference grammars, were written and their publication drove linguistic standardization. In this way, the once non-uniform vernacular languages gradually became unified in orthography and punctuation. While this development presumably applies to the entirety of linguistic communication during the Late Modern Period, the formation of national scientific institutions led to linguistic developments specific to scientific writing due to a sharp increase in scientific publications and academic internal stylistic preferences. For instance, the Royal Society had a clear vision of what scientific language should look like: "They argued that the English prose of scientists should be stripped of ornamentation and emotive language. It should be plain, precise and clear. The style should be non-assertive. Assent was to be gained not by force of words but by force of evidence and reasoning" (Baugh & Cable, 1993, p. 238). While such instructions are rather vague from a modern linguistic point of view and appeal to a certain "style" of what scientific language should look like, they reflect the active promotion of the formation of a linguistically distinctive scientific meta-register (encompassing all kinds of scientific texts) as a result of changing communicative

needs of the scientific community. We can assume that due to a complete turn in terms of scientific methodology and to the enormous increase in inventions and discoveries, the newly arising meta-register was subject to heavy external pressures, such as a significant increase in new incoming vocabulary enormously augmenting lexico-semantic complexity of scientific writing. From a cognitive perspective, such an increase in lexico-semantic complexity can be assumed to pose an important strain on processing.

The question that we ask in the present thesis is thus: how do writers of scientific language maintain *communicative efficiency* in spite of the growing external pressures? Hawkins (1994, 2004, 2014) for instance suggests that grammar plays a regulatory role in the evolution of language use, helping to maintain communicative efficiency, particularly through the variation of word order, taxis, and syntactic embedding serving as means of managing linguistic complexity and reducing the processing effort connected to it. We believe that balancing out lexico-semantic expansion by decreasing lexico-grammatical complexity leads to an optimized code for communication among scientific experts. Coming back to the quote above, we can see that the stylistic instructions fundamentally imply this particular reduction of grammatical complexity by demanding a sharp turn away from linguistic redundancy. This is entirely plausible if we think that in the scientific community, shared expert knowledge grows to such an extent that many explicit grammatical relations over time become obsolete. We can see this development reflected in the natural cycle of the emergence of a new concept: starting with its semantic delineation by explicit linguistic means and ultimately resulting in the creation of a technical term. Take for instance the discovery of chemicals, their presentation to the community, and ultimately the concrete designation of chemical terminology as illustrated in the example of *hydrogen*:

(1) a. *The last, indeed, sufficiently characterizes and distinguishes* ***that kind of air which takes fire****, and explodes on the approach of flame; but it might have been termed fixed with as much propriety as that to which Dr. Black and others have given that denomination, since it is originally part of some solid substance, and exists in an unelastic state, and therefore may be also called factitious.* (Observations on different kinds of air, Joseph Priestley, 1772)

b. *The term mephitic is equally applicable to what is called fixed* ***air, to that which is inflammable****, and to many other kinds; since they are equally noxious when breathed by animals.* (ibid.)

c. *I know of only three metallic substances, namely, zinc, iron, and tin, that generate* ***inflammable air*** *by solution in acids; and those only by solution in the diluted vitriolic acid, or spirit of salt.* (Henry Cavendish, 1766)

d. *After exhausting the air from the jar the* ***hydrogen***[1] *was allowed to pass*

[1] *"inflammable air: This term was applied to hydrogen, H_2, once it was recognized as a distinct air; it was also used as a descriptive term for flammable gases or gas mixtures more generally.*

> *into and through it, and this process was repeated four times.* (W. C. Sturgis, Professor H. Marshall Ward, 1899)

Example (1) shows that when a concept is still new, it is first described using the grammatically highly complex and explicit constructions of *relative clauses* as in Example (1-a), and as the community gains shared knowledge of the concept, encodings gradually become grammatically less explicit by using shorter, more compressed constructions such as attributive adjectives, and finally a new term is born.

In the present thesis, we argue that relative clauses represent especially interesting constructions to trace the counterbalancing of lexico-semantic expansion by a decrease in grammatical complexity. To create a theoretic basis for this claim, in the Background chapter (Part I), we delineate the central theoretic concepts of this thesis – *efficiency*, *utility* and *complexity* – and explain how they are interlinked with each other. Being a rather vague concept, we define what grammatical complexity means in the context of this thesis and identify the different linguistic units in which complexity can be observed *in* and *by means of* relative clauses. To be able to make claims about the degree to which grammar in scientific writing becomes less complex and thus easier to process, we use specific *complexity measures* (Chapter 5) that have empirically been shown to be associated with the cognitive effort involved in sentence processing. This cognitive processing effort can essentially be divided into two types: *memory-based* and *expectation-based* processing effort. Each of the measures can furthermore be associated with a structural linguistic level (i.e. lexis and syntax). For instance, the degree of syntactic complexity of a sentence can be estimated in terms of structural *intricacy*, i.e. "the length and depth of the tactic structures whereby clauses come together to make up a clause complex" (Halliday & Webster, 2004, p. 33). *Intricacy* can thus be modulated by the optional usage of relative clauses as linguistically redundant material (i.e. when other shorter encodings can replace them). We can estimate part of this intricacy in terms of the relative frequencies of relative clauses. Relative clauses themselves can also be constructed in more or less complex ways leading to longer or shorter syntactic dependency relations between the head noun and the embedded verb of the relative clause. This kind of syntactic complexity (known as *locality*) is generally associated with the cognitive processing difficulty for the working memory involved in processing syntactic dependencies (Gibson, 2000). Apart from working memory, expectation also seems to play a crucial role in the way relative clauses are processed. For instance, relative clauses can be modulated in terms of their *accessibility* (Chapter 2.1.3.1), which is connected to the extraction position at which the relativization takes place. According to Keenan & Comrie (1977), relative clauses extracted from the subject position are easier to comprehend than relative clauses extracted from the object position due to the higher expectation of the first.

Beyond the syntactic complexity created by and within relative clauses, they can

[Cavendish, Franklin, Priestley, Watt et al. 1784]" cited from (Giunta, 2023).

also show different degrees of complexity on the lexico-grammatical level. For instance, relative clauses have a wide array of introductory markers constituting the paradigm of relativizers. We call the degree of variability (i.e. number and probability distributions) of the relativizer paradigm *paradigmatic richness*. To estimate the complexity-related processing effort of paradigmatic richness, we use *entropy*, reflecting the uncertainty about an upcoming item, which has been shown to be an indicator of processing difficulty in sentence processing (Genzel & Charniak, 2002). Entropy (Shannon, 1948) is an information-theoretic measure based on probability distributions of different options at a given choice point. Applied to the relativizer paradigm, entropy reflects the uncertainty about the choice of a specific relativizer. If all relativizers have the same probability to occur, then entropy is highest; the more skewed the probabilities are toward one preferred relativizer, the lower the entropy, or uncertainty about the choice of the relativizer.

Also on the lexico-grammatical level, the complexity of relative clauses can be influenced by their syntagmatic predictability in the context they occur in. This means that relative clauses in highly conventionalized contexts (Example (2-a)) are easier to process than those in rather atypical contexts (see Example (2-b)) since in the latter they are much less predictable. We can estimate the processing effort at the point of encountering the relativizer using *surprisal* (based on *Shannon entropy*), another information-theoretic measure that refers to the unexpectedness of an event or message, and represents the bits of information needed to decode the message. The surprisal values in our linguistic corpus data are annotated using a trigram language model trained on different periods of time to capture differences in syntagmatic predictability of items at different points in time.

(2) a. *This secondary transformation depends principally on* ***the manner in*** *which the operation is conducted.* (A. W. Hofmann, 1867)
b. *This secondary transformation depends principally on* ***supplementary processes in*** *which the operation is conducted.*

Our approach to investigating complexity in relative clauses thus encompasses both the lexico-grammatical and the syntactic dimensions of relative clauses. It is based on the assumption that a reduction of complexity on each of the dimensions leads to a reduction in processing effort to counterbalance the pressure deriving from increased lexico-semantic complexity.

The thesis is structured as follows. In the Background part (Chapter 2) of the thesis, we give a general introduction to the concepts (efficiency, complexity, utility) used in this thesis as well as to the historical and linguistic developments involved in the formation of the scientific meta-register in English and German. In Chapter 2.1, we give an overview of the existing literature on communicative efficiency, how it is connected to linguistic complexity, and how both are involved in the formation of the scientific meta-register. Here, we follow the assumption that the means to

create efficiency in a language are to some extent register-specific. Since scientific writing is affected by register-specific pressures, complexity is modulated in such a way as to achieve efficiency for the specific communicative purposes of expert-to-expert communication among scientists. We call this interplay *utility*. We then move on to defining the specific types of complexity we look at in this thesis, dividing them into *lexico-grammatical* and *syntactic complexity*. We explain how they can be assessed with the complexity measures used in our corpus analyses and how they are related to the two types of processing effort (i.e. memory-based and expectation-based). We conclude Chapter 2.1 with a definition of complexity as it is used in the thesis. In Chapter 2.2 we move on to discussing crucial historical developments between 1650 and 1900 which can be assumed to have affected the development of the scientific meta-register in the English- and the German-speaking areas. We believe this to be necessary to build hypotheses about the language-specific developments we expect to encounter in our corpus analyses. In Chapter 2.3, we discuss the state of the art of linguistic changes in the Late Modern Period and the formation of the scientific meta-register in English and German. We first give an overview of the most prominent general changes and then move on to the specific evolution of scientific writing. Since relative clauses are the central subject of our studies, we dedicate a section to extant work on their specific diachronic development as markers of complexity, as well as to the developments regarding their introductory markers during the Late Modern Period. In Chapter 3, we present the hypotheses on the basis of which we conduct our corpus studies connected to the five dimensions of grammatical complexity defined in Section 2.1.3: *paradigmatic richness*, *syntagmatic predictability*, *syntactic intricacy*, *locality* and *accessibility*.

In Part II, we present the corpora we used for our empirical studies. To trace the diachronic development of grammatical complexity in the scientific meta-register, we believe that it is not enough to analyze scientific texts exclusively, but that register-specific developments can be captured much better against the background of an object of comparison. For this reason, for each language, we prepared a scientific corpus and a general language corpus: the *Royal Society Corpus* (RSC) for scientific English and the *Corpus of Late Modern English Texts* (CLMET) for general English, both described in Section 4.1. The German corpora are compiled from texts from the *Deutsches Textarchiv* (DTA): the DTAW contains scientific German texts and the DTAG represents general German; they are described in Section 4.2. As our analyses rely on different types of linguistic annotation, we first introduce the "basic versions" of the corpora including existing linguistic annotation (i.e. lemmas and parts of speech) and surprisal. In Section 4.3, we describe the process of syntactic parsing generating the parsed corpus versions including Universal Dependencies annotations. In Chapter 5, we explain how we calculate the different measures of lexico-grammatical and syntactic complexity based on the different annotations in the corpora. To assess **lexico-grammatical complexity**, we introduce the information-theoretic measures, *entropy* for quantifying *paradigmatic richness* and *surprisal* for assessing *syntagmatic*

predictability, in Section 5.1. We then describe three methods to determine **syntactic complexity** in terms of *intricacy*, *locality* and *accessibility* in Section 5.2.

Our corpus studies are presented in Part III (lexico-grammatical complexity) and Part IV (syntactic complexity). Every study is divided into a macro-analytic part in which we use the complexity measures (presented in Chapter 5) to assess the degree of complexity in each dimension, and a micro-analytic part in which we qualitatively explore the linguistic changes affecting grammatical complexity in each dimension.

Part III consists of the first two corpus studies concerned with *lexico-grammatical complexity*. The first study (Chapter 6) comprises a macro-analysis investigating the development in the *paradigmatic richness* of the relativizer paradigm in scientific and general English and German. To do so, we calculate the entropy of the paradigm in different periods of time. In the second study, the micro-analytic part of the chapter, we inspect the frequency distributions of the different relativizers to explain the encountered entropy trends. In Chapter 7, we investigate the *syntagmatic predictability* of relative clauses over time. For this, we inspect the surprisal of relative clauses given their syntagmatic contexts (lexical trigrams) in general, as well as specific relativizers given their syntagmatic contexts. We qualitatively analyze the most frequent lexical as well as grammatical contexts to discover contexts in which relative clauses become especially conventionalized.

In Part IV, we investigate the development of syntactic complexity created by and reflected in relative clauses. In Chapter 9, we analyze the frequency development of relative clauses in the four corpora to get an understanding of how syntactically intricate in terms of relative clause usage scientific vs. general language has become over time. In Chapter 10, we take a two-step approach. We start with a macro-analytic part measuring the general development of average dependency length in the four corpora. Here we investigate the influence of sentence length and the distributions of short vs. long dependency relations (e.g. those created by relative clauses) on the overall trends of average dependency length. In the micro-analytic part, we investigate the specific development of average dependency length in relative clauses to find out whether they become syntactically less complex over time. In the last chapter of Part IV (Chapter 11), we analyze the overall accessibility of relative clauses in the four corpora over time by looking at the distributions of the different relative clause types.

In Part V, we conclude the thesis with a summary of results by hypotheses, and by stating the implications of our findings for our overarching hypothesis that scientific writing becomes less complex and more efficient over time by adapting to communicative needs posed by the community. We then discuss the limitations and summarize the main contributions of the thesis, and present directions for future work.

Chapter 2

Background

2.1 Efficiency, utility, complexity

The assumption this thesis is built upon is that scientific language as it evolves maintains communicative efficiency despite extra-linguistic pressures which pull in the direction of greater expressiveness, such as the expansion of the specialized scientific vocabulary. In the present chapter, we will start by presenting previous studies on *communicative efficiency* focusing on different kinds of linguistic units (Section 2.1.1). We will argue that on the whole, communicative efficiency is strongly dependent on the respective communicative situation, an idea that is framed in the concept of language *utility*, which we will delineate in Section 2.1.2. Apart from being bound to the communicative situation, efficiency is achieved by the modulation of *complexity* on different linguistic levels. Complexity, however, is a much-debated concept. Thus, in Section 2.1.3, we present the definition of complexity as it pertains to this thesis. In Section 2.1.4, we will come back to the question of how scientific discourse maintains its efficiency over time despite the pressures it faces over the years. The explanation we follow here is that communicative efficiency in scientific language is maintained through a trade-off between different types of complexity in different linguistic units, always in accordance with the communicative functions that scientific language fulfills for its users.

2.1.1 Efficiency and optimal encoding

Efficiency in a language is mostly defined as successful communication with the lowest necessary effort (Levshina, 2018). Gibson et al. (2019, p. 3) specify this even further by assuming that there is a trade-off between successful communication and "minimal effort". Levshina (2018) starts from the assumption that in human communication there is (in most cases) a choice between different linguistic encoding options, and the choice of which option to use is made on the grounds of the principle

of least effort to achieve a certain communicative goal. Theoretical efforts to pin down efficiency have been made for a long time in several linguistic disciplines, i.e. psycho-linguistics, pragmatics, typology, and using different theoretical frameworks (probability and information theory, dependency grammar) and methodological approaches (corpus-based, experimental, computational modeling) to inspect efficiency on different levels of linguistic structure (e.g., phonology, syntax, morphology). Comprehensive overviews are provided by e.g. Gibson et al. (2019) and Levshina (2018). It is generally assumed that languages show a universal tendency toward having an efficient structure. Importantly, Gibson et al. (2019) speak of communicative efficiency and not linguistic efficiency, implying that efficiency is dependent on the participants in communication, which is essentially the purpose of linguistic interaction. The communicative aspect is important, as it includes both the sender and receiver of a message. Communication is efficient if both parties have to invest the lowest possible effort to send and decode a message. Levshina (2018, p. 3) presents an overview of most traditional efficiency theories that employ the principle of least effort, such as Zipf's Principle of Least Effort (Zipf, 1949); the Gricean maxim of Quantity and the Neo-Gricean maxims (Grice, 1975; Horn, 1984; Levinson, 2000); Haiman's (1983) principle of economy; Du Bois' (1985) dictum "Grammars code best what speakers do most"; Keller's (1994, p. 107) hypermaxim "Talk in such a way that you are socially successful, at the lowest possible cost" and maxim "Talk in such a way that you do not spend more energy than you need to attain your goal"; Hawkins' (2014) principle "Minimize Forms"; Givón's (2017, p. 157) code–quantity principle and Haspelmath's (2021) "grammatical form–frequency correspondence hypothesis". Levshina (2018, p. 4) herself then formulates "The Principle of Communicative Efficiency: Communicate in such a way as to maximize the benefit-to-cost ratio" according to which "[t]he communication is efficient when the speaker spends not more and not less energy than it is necessary to cause [the intended] cognitive effects." All the mentioned theories assume rational behavior on the part of the interactants. To test the general human striving for efficiency, information theory (Shannon, 1948) as a theoretical framework especially lends itself to formalizing efficiency in communication. Information theory assumes that the channel of communication contains noise and that a sender of a message formulates her message using a code that is robust to noise, i.e. makes communication possible in spite of the noise (cf. Gibson et al., 2019, p. 3). A channel that is robust to noise and still transmits the message successfully with the lowest possible effort is assumed to be optimal. It is, however, important to note that the amount and quality of noise depend very much on the channel (e.g. written vs. spoken communication) and can in any case only generally be assumed to exist, rather than specifically estimated or measured. Information theory is based on the idea that what is highly predictable needs fewer bits of information for encoding and successful transmission. This assumption can fruitfully be used to explain communicative efficiency, e.g. on the level of lexis, where more frequent and therefore more predictable words are shorter or sometimes even omitted altogether when the information they

convey approaches zero.

Communication is based on the assumption of mutual rationality (Levshina, 2018) according to which the speaker and hearer share the heuristics "low costs – low benefits" (Low-Cost Heuristic) and "high costs – high benefits" (High-Cost Heuristic) (Levshina, 2018, p. 53). When the High-Cost Heuristic is used, the receiver's "previous cognitive state" is subject to a substantial change, while the Low-Cost Heuristic refers to a non-substantial change in the cognitive state of the recipient (cf. Levshina, 2018, p. 5).

Efficiency has also been studied with a view to the diachronic development of languages over time. For instance, Levshina (2018, p. 57) assumes that her Low-Cost Heuristic over time "leads to reduction of forms as an adjustment to the high probability of the information" conveyed by highly recurrent linguistic events, i.e. the more predictable words or expressions become, the shorter they become. This kind of reduction following the Low-Cost Heuristic is called "formal reduction" and can be regarded as a central mechanism of language change (cf. Levshina, 2018, p. 59). In line with this, Langacker (1977) regards "languages in their diachronic aspect as gigantic expression compacting machines" (Langacker, 1977, p. 106, cited after Levshina (2018)). Formal reduction is furthermore bound to the communicative context. For instance, forms can be reduced in situations when the intended receiver of a message can be expected to infer the meaning of a reduced form (cf. Levshina, 2018, p. 60). This statement holds particularly true to scientific language. The participants of scientific communication on a particular subject share a high degree of common knowledge facilitating an efficient style, which can be afforded, since reduced forms can easily be recovered through common background knowledge. The efficient style achieved through this reduction of forms can also be described as a kind of optimal coding for a specific communicative situation. For instance, Degaetano-Ortlieb & Teich (2019) find that scientific language over time develops toward an optimal code. They take an information-theoretic perspective on linguistic change in scientific writing using Kullback-Leibler Divergence (KLD: Kullback & Leibler, 1951) and Surprisal (Shannon, 1948). They follow the assumption that writers of scientific literature are rational and aim to encode their messages optimally by employing particular linguistic choices to regulate the quantity of information conveyed (Degaetano-Ortlieb & Teich, 2019, p. 26). They show that over time scientific writers particularly converge on specific *grammatical* options, which gradually become "conventional[ized] and thus more expected (less surprising)" (Degaetano-Ortlieb & Teich, 2019, p. 26). Conversely, on the lexical level, they find a strong "versatility, which is indexed by informational peaks [indexed by *surprisal*] in phases of lexical innovation/expansion and mid to low levels of information in phases of stability/consolidation" (Degaetano-Ortlieb & Teich, 2019, p. 26). This means that there is a constant intake of new vocabulary leading to highly surprising (unexpected) words in the scientific literature, which over time settle into the vocabulary and become less surprising. In the present thesis, we follow the assumption that this interplay between increasingly efficient (more pre-

dictable) grammatical structures and informationally packed new members in the vocabulary represents a mechanism to counterbalance increasing complexity on the lexico-semantic level by decreasing complexity on the grammatical level.

2.1.2 Utility and register

Having introduced the notion of efficiency and its diachronic implications in the previous section (2.1.1) we would now like to link the notion of efficiency to the particular area of inquiry of this thesis: diachronic change in scientific writing as a meta-register of various registers associated with the domain of science. To do so, we would like to add a register-theoretic notion to efficiency. This thesis is concerned with the early years of modern science and we assume that over time the evolution of the scientific meta-register follows the specific requirements of the user community as it progresses. It is our assumption that register development is a consequence of the aim to continuously adjust language as perfectly as possible to its specific communicative function. To define what the specific communicative function of scientific writing is, it is helpful to break the register configuration up into the three main contextual parameters of a text, namely *field*, *tenor*, and *mode* (Halliday, 1985, p. 12). The *field* of discourse is concerned with the question *what is the nature of the social action that is taking place*. The *tenor* of discourse refers to *the participants of the social action* and the *mode* of discourse describes *the part the language is playing* (Halliday, 1985, p. 12). Obviously over time, as the external conditions of scientific writing change on various levels (society, institutions, technology, etc.), contextual parameters of scientific writing also change. In terms of the *field* of discourse, it is important to note that during the Late Modern English (lModE) period, scientific sub-disciplines (Biology, Chemistry, Physics, etc.) started to develop, making the *field* increasingly diverse and hence requiring more diverse linguistic means of expression, most notably newly created vocabulary in each field. Also, during our time period, scientific research witnessed a remarkable surge, and scientists for the first time started to publish their findings in the vernacular languages (English earlier and to a broader extent than German). The change from writing in Latin to writing in the vernaculars has a non-trivial impact on the *tenor* of discourse, making it possible for people with no command of the Latin language to write and read scientific texts. Intellectual movements such as humanism and the Enlightenment also had a strong impact on the development of science and its linguistic forms of expression in terms of *tenor*. For instance, Enlightenment philosophers promoted the accessibility of knowledge to a wider public, making it necessary to structure language in a clear and comprehensible way. Also, the creation of an increasing scientific community producing an enormous amount of knowledge shared between the community members had a strong impact on forms of expression in scientific writing. Thus, communication between members of a specific community, in terms of *tenor*, became more and more specialized and developed from "expert to layperson" to "expert to expert" communication (cf. Biber & Gray, 2016,

p. 51). The increasing momentum of scientific productivity necessitated an adjustment of the language to the new contexts that scientific communication took place in. Scientific activity was put on a stronger institutional basis by founding scientific institutions like the *Royal Society of London* and by the publication of scientific journals like the *Philosophical Transactions of the Royal Society*, clearly affecting the *mode* of discourse, increasingly shifting towards written modality. While in Section 2.1.1, we discussed the general striving of language users to be efficient in language usage, the mentioned accounts of efficiency are not specific about the influence of situational context on efficiency. The notion of *utility* proposed by Jaeger & Tily (2011) closes this gap, by describing *utility* as "suitability for a certain communicative function" where utility is improved by reducing complexity: "The notion of 'utility' is, however, much broader than processing complexity. Language utility can be understood as relative to a human language user's communicative needs" (Jaeger & Tily, 2011, p. 327). In this thesis, we apply *utility* to the notion of register and define the concept of *utility* as the modulation of complexity according to the specific communicative function of a text. Note that, in this context, complexity can be understood as operating on different linguistic levels and we assume that depending on the communicative function, different types of complexity are at work. More specifically, applied to scientific language, we can assume that the communicative function is that of efficient information transfer between highly specialized experts in a field sharing a high degree of background knowledge. On the one hand, this shared background knowledge makes it possible to cope with the continuous pressure deriving from the intake of new and conceptually complex words to the vocabulary generated by new discoveries. On the other hand, the linguistic code necessarily becomes increasingly complex (harder to process) on the lexico-semantic level. Consequently, in the context of expert-to-expert communication, the use of highly explicit and syntactically complex noun phrase elaborations of the kind in Example (1-a) are rather superfluous and put additional pressure on sentence processing, while complex phrasal constructions of the kind in Examples (1-b-d) are shorter and more efficient, and therefore suitable for the communicative needs of experts. This assumption is based on several studies looking at scientific writing (Halliday, 1988; Halliday & Martin, 1993; Biber, 2006), which attest that on the level of the noun phrase, linguistic complexity shifted from clausal subordination towards non-clausal pre- and post-modification (see Examples (1)). Biber & Gray (2010, 2016) corroborate the findings from these early studies showing that compressed nominal structures, i.e. complex noun phrases (modified with non-clausal components) are distinctive for scientific English, while grammatical explicitness created by elaborated clause structures is not a prominent feature in academic writing, suggesting that a decrease of clausal complexity on the syntactic level improves the *utility* of these constructions for scientific language. Regarding processing, Steels & Beuls (2017) show that structurally less elaborated constructions are often ambiguous. Hence, the resulting ambiguity, "increases complexity in processing. This arises from the fact that the hearer has to eliminate references that may be associ-

ated with particular constructions or interdependences (sic!), between constituents that are not relevant" (Mufwene et al., 2017, p. 19). We assume, however, that in expert-to-expert communication, ambiguity-induced processing effort is compensated for, through background knowledge.

(1) a. *They constitute* ***a new acid, which I purpose to call Evernesic acid****.* (John Stenhouse, 1848)
b. *These analyses give* $C_{18}H_9O_7$*+HO for* ***the rational formula of hydrated evernesic acid****.* (ibid.)
c. *In this plot, the calculation was based upon* ***the first paradigm, i.e. a well-defined scrapie-specific nucleic acid among the heterogeneous background nucleic acids****.* (Kellings et al., 1994)
d. ***The introduction of a small leaden dish of strong sulphuric acid into the case*** *produced the most violent commotion in a film.* (Reinold and Rucker, 1881)

Example (1-a) including a relative clause is syntactically more elaborated (and thus more complex) than the other sentences without clausal subordination. At the same time, Example (1-a) is grammatically much more explicit than the other Examples (1-b-d), defining *who* gave the name of *what* to *whom*. Example (1-a) shows that syntactic complexity created by grammatical explicitness may improve comprehension for a person lacking background knowledge. At the same time, the clausal subordination creates a higher complexity on the sentence level and may be perceived as redundant by experts. Examples (1-b-d), in contrast, represent highly compressed structures, i.e. reducing complexity on the level of the sentence but increasing phrasal complexity. Not only may the phrasal complexity be a factor in increasing comprehension difficulty, but also the highly implicit content of the dense structure may cause a higher processing effort for a non-expert lacking the specialized background knowledge necessary to infer the implicit information, while an expert should have no problem in doing so. Thus, *utility* can be understood as the modulation of different types of complexity suited for a specific situational context, specifically on the level of the *tenor* of discourse.

2.1.3 Complexity

We have taken up the concept of complexity at several points in the previous sections, without, however, providing a comprehensive definition of it. The aim of the present chapter is thus to discuss previous approaches to defining linguistic complexity from the literature and derive from this a definition that will be used in the present work. Linguistic complexity is a much-discussed concept among linguists; however, few works actually define what complexity is (cf. Mufwene et al., 2017, p. 1). Often in linguistics, complexity is defined by the number of parts a (whole) linguistic unit

consists of, and "the more parts the whole consists of, the more complex it is assumed to be, regardless of how the parts interact with each other." (Mufwene et al., 2017, p. 4). This view is for instance maintained by Newmeyer & Preston (2014, p. 182), who state that "the more patterns a linguistic entity contains, the longer its description, and then the greater its complexity" (cited after Mufwene et al., 2017). Along these lines, in many studies investigating linguistic complexity, the number of elements, i.e. the "number of phonemes, morphemes, words, but also relations among variants of such units [...], or yet the number of categories, rules, or constraints that can be posited in a system" is used to approximate its complexity (Mufwene et al., 2017, p. 4). According to this type of definition, morphological complexity is measured in terms of morphemes per word (see Example (2) showing two morphologically highly complex adjectives), and syntactic complexity can be broken down to linguistic units such as phrases, clauses and sentences. On the phrase level, complexity may refer to the length of a phrase (see Example (3) illustrating the intricacy of a noun phrase post-modified by two prepositional phrases) and on the sentence level, complexity may arise from the number of phrases a clause is composed of or the number of clauses a sentence contains (see Example (4) illustrating a sentence with several clausal sub-ordinations).

(2) *Barium-salt of a new acid which I will call* ***tetra-sulpho-di-phenyl-enic*** *acid; the soluble portion contains another new acid, for which the name* ***tri-sulpho-di-phenyl-enic*** *acid may be adopted* (Peter Gries, 1864)

(3) *the erection of a self-recording anemometer on the roof of the Physical Observatory* (Obituary Notices of Fellows Deceased, 1866)

(4) *He tells* ***that*** *all along the Gulf of Persia there are vast numbers of a kind of Locusts,* ***which*** *are edible, and of* ***which*** *our Traveller affirms* ***that*** *he opened one* ***that*** *was six inches long,* ***and*** *found 17 little ones in its belly, all of them stirring.* (Philosophical Transactions, 1676)

This way of looking at complexity, i.e. counting component parts of a unit, was termed "bit complexity" and criticized as "uninformative" by DeGraff (2001, p. 265, cited after Mufwene et al. (2017)). Also, such a fairly simplistic definition of complexity does not take into account the relationships among component parts in such a complex system as language. Thus, a recurring theme in attempts to define complexity is that of "the coexistence of components that interact with each other" (Mufwene et al., 2017, p. 2). As an attempt at defining and measuring morphological complexity, Nichols (1992), for instance, develops "a measure of complexity based on the number of points at which a typical sentence is capable of receiving inflection" (Juola, 1998, p. 2). For syntactic complexity, for instance, Biber & Clark (2002) suggest that the complexity of a noun phrase can take two forms, that of *compression* and that of *elaboration*. Phrasal compression is created by non-clausal pre- and postmodification strategies, while elaboration refers to modification by means of clausal postmodifiers.

Both forms can be regarded as two distinct *types of complexity*, termed as *clausal and phrasal complexity* by Biber & Gray (2016) illustrated in Figure 2.1, where clausal complexity on the level of the noun phrase corresponds to elaboration and phrasal complexity corresponds to compression.

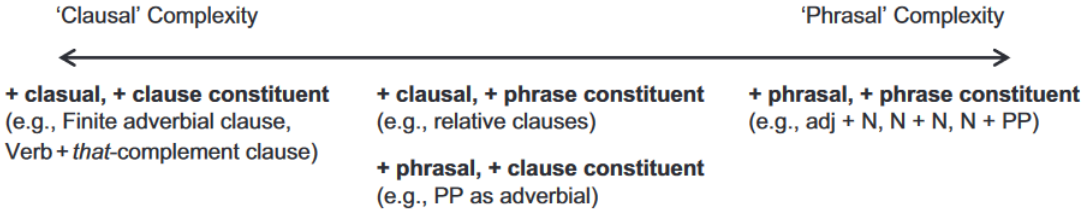


Figure 2.1: A cline of grammatical complexity (Biber & Gray, 2016, p. 62).

While such definitions of complexity may be intuitive and plausible, they are relatively vague in terms of how the degree of complexity can be measured. Apart from this, a definition of complexity exclusively looking at linguistic structure ignores the actual motivation to investigate linguistic complexity at all, namely answering the question "which linguistic structures are difficult to process and why?". Thus, apart from defining complexity on the level of linguistic structure, a cognitively motivated definition of complexity seems better suited to explain the mechanisms underlying communicative efficiency. Indeed, complexity is often associated with *cognitive processing effort*, "the cognitive efforts required by the mind (conceived of as the state of the brain in activity) to produce or to process a message" (Mufwene et al., 2017, p. 11). Just as complexity can manifest itself in different linguistic units (words, phrases, sentences, etc.), processing effort has also been shown to be correlated with the degree of complexity of a specific linguistic unit. Processing effort can be attributed to either a person's **working memory** (how long we have to keep a linguistic unit in mind during online processing) or to a person's **expectation** about an upcoming linguistic event (e.g. which word comes next). Each of these processing principles can be directly related to linguistic phenomena on the two structural levels *lexis* and *syntax*.

In the following, we will first review extant approaches investigating the connection between lexical (Section 2.1.3.1) and syntactic (Section 2.1.3.2) complexity on the one hand, and processing effort on the other. Since this thesis concentrates on the syntactic construction of relative clauses (RCs), we determine specific types of lexical and syntactic complexity detectable in RCs and how these types of complexity are connected to expectation- and memory-based processing effort. On the basis of this, in Section 2.1.3.3 we establish the complexity definition that we follow in the analyses of this thesis.

2.1.3.1 Lexical complexity

Lexical complexity, i.e. the complexity of a word, can be understood merely in a structural way as the number of components (e.g. syllables, morphemes, etc.) the word is composed of. Beyond this, there are various alternative ways of looking at processing-related lexical complexity. For instance, a great amount of research from psycho-linguistics looks at lexical complexity from a semantic point of view (e.g. Kintsch, 1974; Cutler et al., 1983) and frames it as "the relative complexity of meaning representations of a lexical item" (Rayner & Duffy, 1986), i.e. the number of *semantic components* of a word. For instance, causative verbs such as *convince* ("cause to believe") have two components and are considered to be more complex than verbs with only one semantic component such as *sleep* (cf. Rayner & Duffy, 1986). Another factor of semantic lexical complexity is lexical *ambiguity*. However, in studies testing for a relation between this type of complexity and processing demand, no correlation was found (Kintsch, 1974; Cutler et al., 1983; Rayner & Duffy, 1986). Further studies on semantic complexity are reviewed by Jaeger & Tily (2011, p. 325), such as "conceptual accessibility" (Bock & Warren, 1985), which has also been shown to contribute to processing difficulty (in production and comprehension), as well as *imageability* (Bock & Warren, 1985), *prototypicality* (Kelly et al., 1986; Onishi et al., 2008), *animacy/humanness* (Bresnan et al., 2007), *givenness* (Bock & Irwin, 1980) and "semantic similarity to recently mentioned words" (Bock, 1986). These studies show that semantics has an important effect on sentence processing as well. However, in the present thesis, the focus is on RCs representing grammatical structures rather than on the semantic side. A more theory-neutral, probabilistic measure of lexical complexity is *word frequency* being strongly correlated with processing effort: highly frequent words have shown to be easier to process (as tested by word fixation times) and less frequent words are harder to process (e.g. Rayner, 1977; Rayner & Duffy, 1986). Also, *word length* has been mentioned as a factor correlated with reading times (Just & Carpenter, 1980; Rayner, 1977). However, the word length is itself strongly correlated with frequency (compare Zipf's Law, Zipf, 1949). Rayner & Duffy (1986) control for word length and still find a strong effect of word frequency on fixation times.

While all these measures are plausible and very possibly involved in human mechanisms of language processing to some degree, they ignore one crucial factor: ***ambient context***. Since language is rarely processed as isolated words, it seems obvious that what we conceive as being complex is the result of an interaction between the linguistic unit and its context. For this reason, the type of complexity we look at in this thesis is a subarea of lexical complexity, namely ***lexico-grammatical complexity***, by which we mean the complexity of words in their ambient contexts. Ambient context can be defined on two dimensions: the *syntagmatic* dimension referring to the left and right context of a word in a sentence, and the *paradigmatic* dimension considering an array of alternative linguistic options at a given choice point in language produc-

tion and prediction. The idea of considering the context of a word is taken up by information-theoretic approaches to language processing (Shannon, 1948). In information theory, the information content conveyed by a word based on its predictability in its syntagmatic context can be calculated in terms of *surprisal*. Depending on the variability of the contexts that a word can occur in, we assume that the degree of lexico-grammatical complexity of a word is modulated by its *syntagmatic predictability*. Another information-theoretic concept describing the uncertainty about a word to be chosen out of a set of options at a specific choice point is *entropy*. A set of options at a specific choice point in a sentence (e.g. the onset of an RC) can, for instance, be a paradigm (e.g. the relativizer paradigm). The uncertainty about which one of the options will be chosen at a particular point is determined by the number and probability distribution of the different members of the paradigm. We therefore call the type of lexico-grammatical complexity determined by the paradigmatic context of a word *paradigmatic richness*.

Both *syntagmatic predictability* and *paradigmatic richness* are related to the expectation-based strand of processing theories. In the following, we will present the two concepts in more detail and discuss previous work on the two information-theoretic measures (*entropy* and *surprisal*) associated with them.

2.1.3.1.1 Syntagmatic predictability and surprisal Mathematically, surprisal is calculated as the negative *log* probability of a linguistic unit (e.g. word) given a certain amount of preceding linguistic units (e.g. lexical n-grams) and gives the amount of *bits of information* carried by the linguistic unit in its context. In simple terms, surprisal describes the unexpectedness of a linguistic item occuring in a specific syntagmatic context. The unit in which surprisal is measured is *bits of information* Shannon (1948) quantifying the information content of an item, i.e. how much you learn from a particular piece of information. The crucial link between lexico-grammatical complexity and surprisal is that the degree of (un)expectedness of a word in a context depends on the degree of variability of the available contexts that a word can occur in. To illustrate the syntagmatic predictability of a relativizer (e.g. *which*), let us take a look at two Examples of relativizers preceded by three words (a 3-gram):

(5) ... by means of *which* [...]
... the integrity of *which* [...]

Comparing Example (5-a) to (5-b), *which* is much more predictable given *by means of* and much less predictable and thus more surprising given *the integrity of*. This is due to the fact that in language use, *which* generally occurs much more often in the context *by means of* than in the context *the integrity of*. Thus, the higher the probability of a given word in a particular context, the fewer bits are needed to encode it and the less surprising is its occurrence in this particular context. A word with a

very low probability in a certain context requires more bits to be encoded and is thus more surprising. The degree of predictability depends on the particular configuration of the different available contexts available for a specific word. It is thus plausible that in a corpus, (5-a) would occur more frequently than (5-b), and thus *which* in the context of (5-a) would have a higher predictability than in the context of (5-b). For an example of calculation see Section 5.1.2. Applied to the diachronic perspective in this thesis, the distributions of the contexts of relativizers can be assumed to have changed over time and surprisal can thus help to assess the syntagmatic predictability of relativizers at different points in time.

Being an indicator of expectation-based processing effort, surprisal seems like an adequate measure to estimate the lexico-grammatical complexity of relativizers in their syntagmatic contexts over time. Behavioral approaches studying the connection between surprisal and language processing have shown that surprisal is positively correlated with processing effort. For instance, studies using response time measures have shown that participants take longer to process sentences with more surprising words or syntactic structures (e.g. Hale, 2001). Levy (2008) and Smith & Levy (2013) use self-paced reading tasks to measure reading times for sentences with varying levels of syntactic complexity. They find that surprisal is a significant predictor of reading times, even after controlling for other factors known to affect reading difficulty, such as word length and frequency. Using eye-tracking techniques, Demberg & Keller (2008) show that the processing difficulty of a sentence is affected not only by the surprisal of individual words but also by the surprisal of the sentence as a whole. Smith & Levy (2013) find that surprisal is a significant predictor of reading times, even after controlling for other variables like word frequency and length.

Apart from behavioral studies, the connection between surprisal and language processing has also been studied using neuroimaging techniques. For lexical surprisal, Frank et al. (2013) use EEG to investigate the effect of surprisal on brain activity as indicated by the amplitude of the N400 (a negative electrical signal recorded from the brain). They find that surprisal is a significant predictor of the amplitude of the N400 and can, therefore, be regarded as a "generally applicable measure of processing difficulty during language comprehension" Frank et al. (2013, p. 1). Investigating syntactic surprisal (i.e. the surprisal of a syntactic category), Henderson et al. (2016) use functional magnetic resonance imaging (fMRI) to trace the effect of words with high and low syntactic surprisal on sentence processing and find that less expected items elicit higher neural activity than more expected ones. Expectation thus seems to play a crucial role in what we perceive as complex, and surprisal has been shown to be a reliable measure to estimate processing-related complexity on the lexico-grammatical level.

2.1.3.1.2 Paradigmatic richness and entropy Like surprisal, entropy is another measure of information content in a linguistic system; however, the two measures differ in their focus and application. Entropy is a measure of the overall unpredictabil-

ity of a group of linguistic items, such as members of a paradigm. Entropy provides a way to quantify the diversity and complexity of the paradigm by calculating the average amount of information associated with each unit. Mathematically, this is done using Shannon's entropy formula (Shannon, 1948), which takes into account the frequency and probability of each member belonging to the paradigm; for an example calculation see Section 6.2. The resulting entropy value provides a quantitative measure of the complexity and diversity of the paradigm, which we call *paradigmatic richness*. Higher entropy values indicate a greater *paradigmatic richness* and a greater degree of unpredictability of each member of the paradigm, while lower entropy values indicate a lower *paradigmatic richness* and thus a greater degree of predictability.

Applied to the relativizer paradigm, entropy describes the level of uncertainty with which a specific relativizer might occur (or the difficulty of guessing that exactly this relativizer will appear). From this it follows that entropy is highest when all possible relativizers available at a given point of choice have the same probability. Entropy decreases the more skewed the probability distribution is towards one specific relativizer as compared to the alternative options, and is lowest (equal to 0) when one of the relativizers has a probability of 1, i.e. it is "deterministically known in advance" (Genzel & Charniak, 2002, p. 1).

The degree of uncertainty, i.e. the number of bits of information as calculated by entropy, has been proposed as a measure of linguistic complexity influencing processing effort in language production and comprehension. For instance, entropy has been used to model morphological processing by calculating the entropy of inflectional paradigms. Moscoso del Prado Martín et al. (2004) use paradigmatic entropy to determine the effect on morphological processing as measured by response latencies (the time between a stimulus and a reaction). They show that entropy as a measure to predict response latencies is superior to traditional type-token-based counts. They look at inflectional paradigms, which they define as "a random variable whose possible values are the different inflected forms that a base word can take" (Moscoso del Prado Martín et al., 2004, p. 6). To calculate the entropy of an inflectional paradigm, they determine the probabilities of each of the paradigm's members (the different inflectional forms). In a similar vein, Milin et al. (2009) use entropy to measure processing load related to inflectional paradigmatic relations. Milin et al. (2009, p. 2) calculate "the amount of information carried by [an inflectional] paradigm" (all possible inflectional options of a lexeme) by determining the distribution of probabilities of each inflectional option. To model the uncertainty about which specific relativizer will occur at the onset of an RC in a particular period of time, as intended in the present work, entropy can be calculated over the probability distributions of the different members of the paradigm of relativizers at different points in time. The more skewed the probabilities are towards more probable options, the lower the processing effort at the choice point of a relativizer.

2.1.3.2 Syntactic complexity

Having reviewed extant approaches to tackle processing-related lexico-grammatical complexity, let us now look at ways that have been proposed to approach syntactic complexity. Biber & Gray (2016)'s concepts of syntactic complexity distinguish between *phrasal complexity* created by means of syntactic compression (e.g. by using attributive adjectives to modify a noun phrase) and *clausal complexity* achieved through the elaboration by clausal subordination (e.g. by using relative clauses to elaborate a noun phrase). The distinction between clausal and phrasal complexity has been extensively investigated especially with regard to academic writing (e.g. Gray, 2015; Staples et al., 2016). The studies have shown that in present-day academic English, phrasal complexity is much more characteristic than clausal complexity, and that the preference for phrasal over clausal complexity has gradually developed over time (Biber & Gray, 2016) as a register feature of academic writing.

To explain these observations, the notion of *utility* seems like a promising concept to consider. As Mufwene et al. (2017, p. 5) notes, "the descriptive account of complexity can be at odds with a more functional approach". This means that the complexity at one structural level may seem high, while the perceived processing difficulty may actually be relatively low. Hawkins (2009) also refers to this apparent contradiction by noting that some longer utterances may sometimes be easier to process than shorter encodings, and subsumes this phenomenon under the term "effective complexity"; Mufwene et al. (2017) lays out the possibility that "the lesser complexity of a module [i.e. linguistic unit] will be balanced by the greater complexity of another." (Mufwene et al., 2017, p. 13 f.). Take for instance the following two noun phrases (Example (6)), the first being an original title of a scientific article and the second being its explicitated translation[1] to a register understandable for non-experts.

(6) a. *Microplastic ingestion by riverine macroinvertebrates* (Windsor et al., 2019)

b. *Ingestion of fragments of any type of plastic which are less than 5 mm (0.20 in) in length by animals big enough for us to see without using a microscope that neither possess nor develop a backbone and which live in rivers.*

Example (6-a) may be harder to process for the non-expert reader due to a lack of background knowledge necessary to infer implicit relations. Example (6-b) is much more explicit, but working memory is strongly burdened by the long-distance dependency relations. In line with the statement by Edmonds (1995) "that complexity lies before all in the eye of the interpreter of the system" (cited after Mufwene et al., 2017, p. 5), our assumption is that one version can be more or less complex and thus efficient depending on the communicative situation and the background knowledge of the reader. Applied to the concepts of phrasal and clausal complexity, it seems plau-

[1] Definitions taken from (Utah State University, 2020).

sible that scientific writing has become less complex in terms of clausal complexity for the purpose of greater *utility* in the specific communicative situation of scientific writing.

The use of clausal complexity in terms of subordination generally leads to stronger ***syntactic intricacy***: "the length and depth of the tactic structures whereby clauses come together to make up a clause complex" (Halliday & Webster, 2004, p. 33). Applied to relative clauses (RCs), the "length and depth of the tactic structures" (ibid.) creates long-distance dependencies between the syntactic head, i.e. the head noun that is being elaborated on, and the predicate of the RC. Since the resulting cross-clausal dependencies require the storage of information in working memory across a prolonged time span, we assume that intricacy can be regarded as a memory-related complexity type. However, the degree of clausal complexity (which we here call *syntactic intricacy*) has mostly been measured in terms of simple frequencies of subordinate clauses and has not been directly associated with cognitive processing effort. It can nonetheless be assumed that the frequency of RCs within and across sentences has an influence on the overall processing difficulty associated with a text.

Another, more processing-related, type of syntactic complexity is ***locality***, which refers to the principle that the relationship between two linguistic elements in a sentence depends on their proximity to each other in the sentence. The degree of *locality* can be measured by the linear distance between a syntactic head and its dependent, i.e. *dependency length* (DL). DL has been shown to correlate with *memory-based* processing effort. Apart from memory-related complexity, specifically, in relative clauses (RCs), syntactic complexity can also become manifest in *expectation-based* processing effort depending on the ***accessibility*** of the RC (i.e. the syntactic position that an RC is extracted from). Thus, syntactic complexity can be created and modulated by the frequency of RCs in the form of syntactic intricacy, by dependency length, as well as the choice of the extraction type of an RC. In the following, we will review relevant literature focusing on (a) *locality*, i.e. *memory-based* syntactic complexity modulated by dependency length, and (b) *accessibility*, i.e. *expectation-based* syntactic complexity modulated by the extraction type of an RC.

2.1.3.2.1 Locality and dependency length Broadly speaking, syntax describes the compositional combinations of words forming sentences (cf. Gibson et al., 2019, p. 8). When further analyzing the composition of the different functional elements of a sentence, we can do so by analyzing *syntactic dependencies* existing between syntactic *heads* (e.g. *flux* in Figure 2.2) and their *dependents* (e.g. *used* in Figure 2.2), the latter being the elements further defining the *heads*. The "most general notation for describing syntactic dependencies among words is called dependency syntax" (Gibson et al., 2019, p. 8, citing Hudson (1991)).

Syntactic *locality* has been proposed as a concept to determine the complexity of syntactic dependencies since it pertains to the idea that the connection between two language components within a sentence is determined by their linear distance from

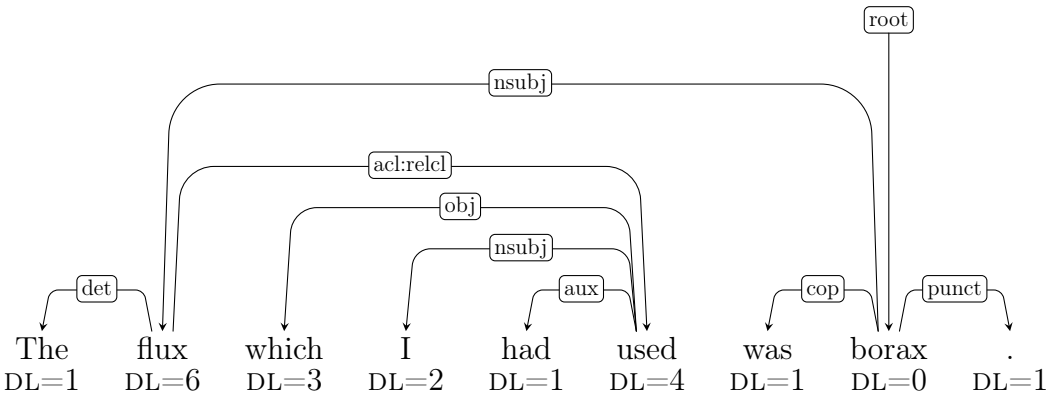
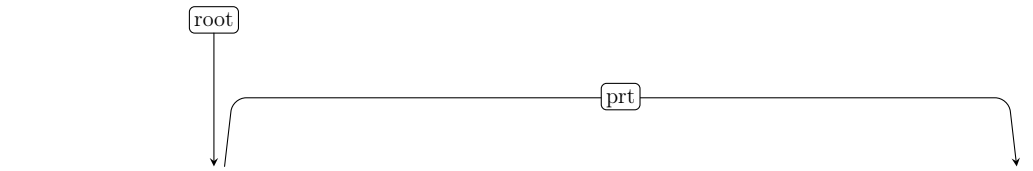


Figure 2.2: Example RC annotated with Universal Dependencies (labels of the edges) and dependency length (DL, labels on the nodes).

each other within the sentence. The linear distance in tokens between a syntactic head and its dependent (cf. Futrell et al., 2015, 2020; Gibson et al., 2019) was originally termed by Heringer et al. (1980) as *dependency length* (DL). *Locality* and *distance* can thus be understood as two sides of the same coin. For instance, in our Example (Figure 2.2), *used* is four tokens away from its head *flux*. DL as a metric of memory-based syntactic complexity derives from the idea prominent in psycholinguistics that "long dependencies correspond to human parsing difficulty due to working memory pressures" (Futrell et al., 2020, p. 375). This idea is framed by *Dependency Locality Theory* (DLT: Gibson, 1998, 2000). According to DLT, the higher processing effort associated with longer dependencies is "time-invoked" (Liu et al., 2017), in that the further away from each other two constituents are, the longer the first constituent must be stored in memory until integration with the second constituent, leading to an increase in processing effort. The resulting "processing slowdown" (Futrell et al., 2020, p. 375) has been proven in experimental setups by longer reading times (e.g., Grodner & Gibson, 2005) as well as lower "speed and accuracy" of the dependency resolution (Nicenboim et al., 2015, p. 2) for sentences with longer DL. As a result, sentences with longer DL are assumed to be less favorable for human language processing than sentences with shorter DL. In this line of thinking, long-distance structures are regarded as more complex than structures with shorter distances. The analysis of particle verbs illustrates this point: In Figure 2.3a, the particle *over* has a long dependency length to its head *brought* and the sequence is relatively hard to process. This contrasts with Figure 2.4b, where the dependency length between the particle and its head is short, being more favorable in terms of working-memory load. Thus, the human processor will prefer the sentence with the shorter DL.

This idea is framed in the *dependency length minimization* (DLM) hypothesis. DLM is a well-established principle representing the assumption that universally, word orders coincide in their tendency to minimize the distance between two closely related words (e.g. Hawkins, 1994; Rijkhoff, 1990; Wasow, 2002; Ferrer i Cancho, 2004; Liu, 2008; Gildea & Temperley, 2010; Futrell et al., 2015), assuming that minimizing this



(a)

(b)

Figure 2.3: (a) Discontinuous particle verb. (b) Continuous particle verb.

distance results in lower processing effort in both language production (e.g. Hawkins, 1994, 2004) and comprehension (Gibson, 1998, 2000). Though DLM is widely acknowledged to be a language universal explaining the human tendency to communicate efficiently by building sentences that are "easy to produce and comprehend" (Futrell et al., 2015), the underlying theoretical frameworks, such as Phrase Structure Grammar (Hawkins, 1994; Gibson, 1998) and Dependency Grammar (Hudson, 1995), methodological approaches (experimental and corpus-based), and objects of research working with DLM are quite diverse. The diversity of approaches entails various, if only slight, differences in the way dependency length (also called dependency distance: Liu, 2008; Liu et al., 2017) is calculated and the methods of testing its validity. For instance, in the past 20+ years, there have been numerous corpus-based studies testing for DLM's universality by comparing the DL of naturally occurring word orders to that in random orderings. In all studies it was found that DL in natural word order is minimized as compared to a random baseline. The results were obtained in studies focusing on a single language (e.g. for Chinese, Liu, 2008) or in contrastive studies, e.g. of Czech vs. Romanian (Ferrer i Cancho, 2004) or English vs. German (Gildea & Temperley, 2010), as well as in large-scale studies comparing 20+ different languages (Liu, 2008; Futrell et al., 2015, 2020) annotated with Universal Dependencies (UD). Although differing with regard to the number of languages observed, the way dependency length (or distance) is calculated, and the methods of creating random baseline models, all the studies coincide in concluding that DL is minimized in natural language as compared to artificial random orderings.

Apart from the corpus-based studies mentioned above, there have been numerous psycho-linguistic experimental studies focusing on the connection between the processing effort connected to specific syntactic constructions and the DL created by the constructions. Here, we review those studies focusing on RCs representing a syntactic pattern variable in terms of its syntactic point of extraction (i.e., subject vs. object)

and thus (among other factors) modulating the DL between the head noun and the embedded verb.

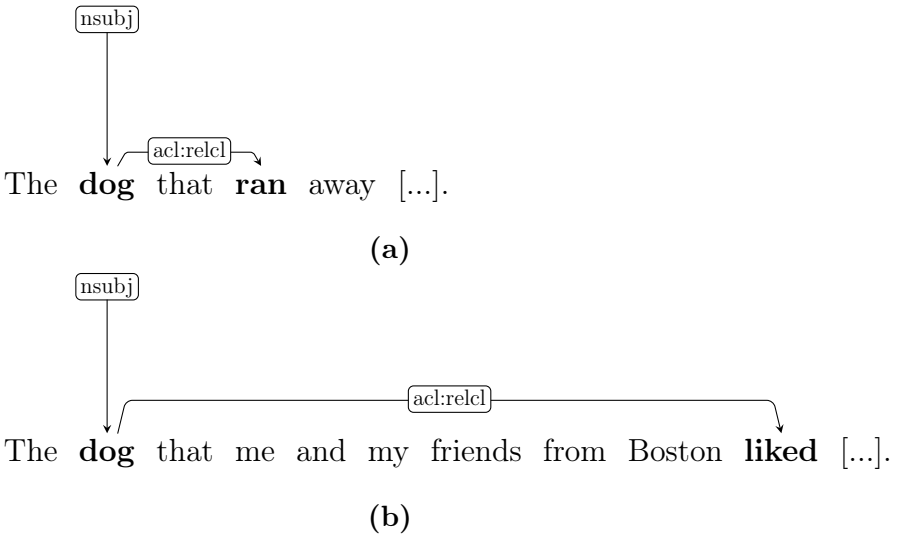


Figure 2.4: (a) Subject RC. (b) Object RC.

Early experimental insights into processing effort modulated by DL came from experiments testing comprehension difficulty of different types of relative clauses (RCs) in different languages (Gibson, 1998; Hsiao & Gibson, 2003; Grodner & Gibson, 2005; Levy & Keller, 2013; Gibson & Wu, 2013). Based on the *Accessibility Hierarchy* or AH (Keenan & Comrie, 1977; Lehmann, 1984, for English and for German, respectively; see Equation 2.1), the processing effort of RCs depends on their syntactic extraction position, i.e. subject vs. direct object, etc. The hierarchy predicts that subject RCs (for an illustration, see Figure 2.4a) are the easiest to process, followed by direct object RCs (Figure 2.4b), etc. For instance, Gibson (1998) finds that comprehension of subject RCs is easier than that of object RCs in English, attributing the processing advantage to locality, i.e. the shorter DL between the head noun and embedded verb in subject RCs. For Chinese, comprehension tasks suggested the same locality effects with shorter reading times (RTs) for object RCs, since in Chinese the distance from the RC verb to the object RC's head noun is shorter (Hsiao & Gibson, 2003; Chen et al., 2008; Lin & Garnsey, 2011). Strong locality effects were also found in Levy & Keller (2013) for Russian RCs. Beyond RCs, further studies on other linguistic phenomena (e.g. Bartek et al., 2011, looking at subject-verb dependencies in English) suggest the universality of increased processing cost due to increased dependency length (Fedorenko et al., 2013; Levy & Keller, 2013; Hofmeister et al., 2007; Hofmeister & Sag, 2010).

2.1.3.2.2 Accessibility In the previous Section, we have presented the subject advantage as the most efficient type of RC in terms of DL and thus working memory load. Alternatively, the subject advantage can also be explained in terms of

"expectation-based accounts" (Chen et al., 2021). Expectation-based accounts (e.g. Roland et al., 2007) trace the subject advantage back to the higher frequency with which subject RCs (Example (7-a)) occur compared to other extraction types (Examples (7-b)–(7-d)).

(7) a. *The reaction which converts Y to Z...* (Subject RC)
b. *The reaction which we describe in the article...* (Direct Object RC)
c. *The reaction which we give the name Reimer Tiemann reaction...* (Indirect Object RC)
d. *The reaction with which we aim to trigger...* (Oblique RC)

Since subject RCs are most frequent (followed by direct object RCs, etc.), the order of accessibility follows distributional probabilities and can thus be seen as expectation-based when it comes to processing the constructions. In English and German, RCs can be extracted from "almost any major NP position in simplex sentences" (cf. Keenan & Comrie, 1977).[2]

$$SUBJ. \supset DOBJ. \supset IOBJ. \supset OBL. \supset GENITIVE \supset OCOMP \tag{2.1}$$

The accessibility hierarchy (AH) has two main components. First, it states that languages vary "with respect to which NP positions can be relativized and that the variation is not random" (Keenan & Comrie, 1977, p. 66). Non-random means that the possibility of whether an NP position can be relativized is ordered in terms of a hierarchy expressing "the relative accessibility to the relativization of NP positions in simplex main clauses" (Keenan & Comrie, 1977, p. 66): If a language can relativize from an NP position, it can also relativize from any other position to the left of it in the hierarchy, but not necessarily to the right of it. The hierarchical order of extraction positions is assumed to be tightly connected to the processing difficulty of an RC, with the subject-extracted RC being the easiest to process (Biber et al., 1999). This assumption is held to be true for virtually all Indo-European languages and has been attested for diverse manifestations of processing difficulty such as "lower accuracy, longer processing time, more working memory burden" in comprehension, "slower production, slower responses, more errors, more omissions/substitutions" in production, and "later emersion and acquisition, [and] more avoidance" in L1 and

[2]Note that in German it is not possible to relativize from the position of an object of comparison:

(i) **Der junge Mann, als der Mary größer ist.*

while in English the following rendering would be acceptable:

(ii) *The young man who Mary is taller than.*

L2 language learning (Lau & Tanaka, 2021). The proposed reasons for the subject/object asymmetry are manifold. For instance, *canonicity* has been mentioned as a possible source of difficulty, assuming that a receiver by default expects and thus prefers canonical word order when processing a sentence (Love & Swinney, 1998; Sekerina, 2003). The canonicity of English subject-verb-object (SVO) word order is violated in all RC types to the right of subject RCs (compare the object RC in Figure 2.4b with SOV order). However, the canonicity explanation does not account for the subject advantage in German RCs. Due to the verb-last ordering in German subordinate clauses, subject-object-verb (SOV) word order is identical in all German RC types. Another explanation of the subject advantage is connected to frequency, assuming that more frequent constructions reflect readers' expectations and are therefore easier to process (cf. Ambridge et al., 2015). Since subject RCs are more frequent in almost all languages (Chen et al., 2021, citing Roland et al. (2007)) readers are more used to reading subject RCs than other types of RCs and thus have a higher expectation for encountering a subject RC. Accessibility of RCs has also been looked at from an information-theoretic point of view, such as the Entropy Reduction Hypothesis (ERH: Hale, 2003) stating that the more informative the input (i.e. word), the lower the entropy of the upcoming word. Hale (2006) applies this principle to the processing of relative clauses and uses entropy to calculate the "uncertainty about the rest of a sentence". He tests the ERH on the predictions of the AH and finds that processing of RCs is harder the further down the AH the RC is extracted from. Entropy reduction (ER) is calculated by building a Minimal Grammar and turned into probabilistic versions by using the relative frequency estimation technique (Chi, 1999). The probability weights are assigned according to corpus data from the Brown Family of Corpora (Kučera & Francis, 1967). The outcome supports the experiments by Keenan & Hawkins (1987) showing that processing of lower AH extractions is harder than for higher AH extracted positions. It is, however, important to note that the outcome is tightly connected to the probabilities obtained from the specific case of the Brown corpora. The probability distributions of the different RC types may be different in other linguistic varieties (e.g. registers) such as scientific discourse.

More fine-grained distribution-based studies consider *animacy* of the RC head noun, showing this to be a strong indicator of RC type: subject RCs more frequently have an animate head, while object RCs tend to have an inanimate head (Roland et al., 2007). Also, *pronominality* seems to be predictive of the RC type, i.e. object RCs frequently occur with a pronominal head and are therefore easy to process (Reali & Christiansen, 2007).

2.1.3.3 Definition of complexity in this thesis

In the previous sections of this chapter, we have discussed different types of complexity, which manifest themselves on different linguistic levels, and which are related to processing. We would now like to establish the definition of complexity that we

will be using in the present thesis. The kind of complexity we are after to trace diachronic change in scientific language is inherently *processing-related* and it describes the variability of different relationships existing between the components of linguistic units in relative clauses (as visualized in Figure 2.5) both on the lexico-grammatical and the syntactic level. For this reason, we distinguish between *lexico-grammatical complexity* and *syntactic complexity*. Our definition of ***lexico-grammatical complexity*** looks beyond the internal complexity of a word (number of morphemes) and takes into account the syntagmatic *and* the paradigmatic contexts of a word: more precisely, a word's frequency and its distribution in different *syntagmatic* (2.5, blue circled) and *paradigmatic* (2.5, yellow circled) contexts. The variability of these contexts affects the probability of a word occurring in one of these contexts, which is why our definition of lexico-grammatical complexity is expectation-based. Applied to the specific case of relativizers, the *syntagmatic variability* refers to the left (trigram) context in which relativizers as introductory markers of relative clauses may occur (see Example (8)).

(8) a. *[...] some such stone as the* ***Asyctos of Pliny which*** *once heated will hold so for a week [...]*
b. *I therefore made divers trials,* ***some of which*** *did not displease me [...].*

The sample sentences in Examples (8) show two possible contexts that *which* may appear in. Intuition suggests that *which* in Example (7-a) is harder to predict than *which* in Example (7-b). Thus, apart from the general frequency of a word, the more variable the left context of a word is, the harder it is to predict this specific word from context; we therefore call this indicator of lexico-grammatical complexity *syntagmatic predictability*, and we calculate it in terms of *surprisal* (for a detailed description see Section 5.1.2).

On the paradigmatic axis, lexico-grammatical complexity is modulated according to the variability in choice amongst (i.e. the number of) the different members of a paradigm (e.g. that of relativizers) and their probability distributions. Thus, the paradigmatic context of relativizers refers to the different options available in the same word class of relativizers (*which*, *that*, *who*, *whose*, etc.; see Example (9)). We will therefore call this indicator of lexico-grammatical complexity *paradigmatic richness*.

(9) a. *In the mean while he gathered a subcutaneous Water, of* ***which*** *yet he was afterwards well cured.*
b. *In the mean while he gathered a subcutaneous Water,* ***that*** *[...] of.*
c. *In the mean while he gathered a subcutaneous Water,* ***whereof*** *[...].*

In the sample sentences in Examples (9), the different relativizers are interchangeable. However, the probability of each of the relativizers occurring may be different. Let us assume the following probabilities of the different relativizers within their respective paradigm: (9-a) 0.5, (9-b) 0.3, and (9-c) 0.2. The higher the probability of one

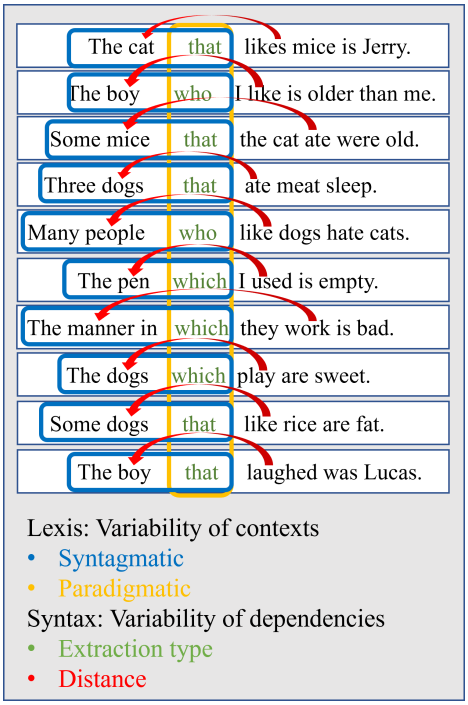


Figure 2.5: Manifestations of *lexico-grammatical* and *syntactic* complexity in relative clauses.

compared to the other members of the paradigm, the lower the uncertainty about the word to occur and the easier it is to predict compared to the other members in the paradigm. We estimate this uncertainty in terms of *entropy* (for an example of calculation see Section 5.1.1). Thus, in our definition, lexico-grammatical complexity can be determined by the variability of syntagmatic as well as paradigmatic contexts and affects expectation-based processing cost.

Our definition of ***syntactic complexity*** revolves around the structural level of syntactic dependencies and includes, but is not limited to, the number of elements (e.g. sub-clauses) in a unit (e.g. a sentence). Taking up Biber and Gray's (2016) notion of clausal and phrasal complexity, we consider *syntactic intricacy* as an indicator of syntactic complexity, created by the presence of sub-clauses (such as RCs) within a sentence. Here, we consider overall RC frequencies as well as their accumulative occurrence within a sentence. Beyond the *number* of specific dependencies contributing to complexity, in our definition of complexity, *locality* is a major determiner of *memory-based* syntactic complexity and is indicated by the dependency length (DL) of syntactic dependencies holding between a syntactic head and its dependent. In our definition, the third indicator of syntactic complexity specific to RCs is *accessibility* (Keenan & Comrie, 1977), that is, the syntactic position an RC is extracted from (i.e. subject RC, direct object RC, etc.; see Example (10)), which has been shown to be mostly related to expectation-based processing effort.

(10) a. *Behold the Figure well, which represents this little Engine in its natural*

size. (Subject RC, Philosophical Transactions, 1672)

b. *Behold this little Engine well, which the Figure represents in its natural size.* (Object RC, generated alternative)

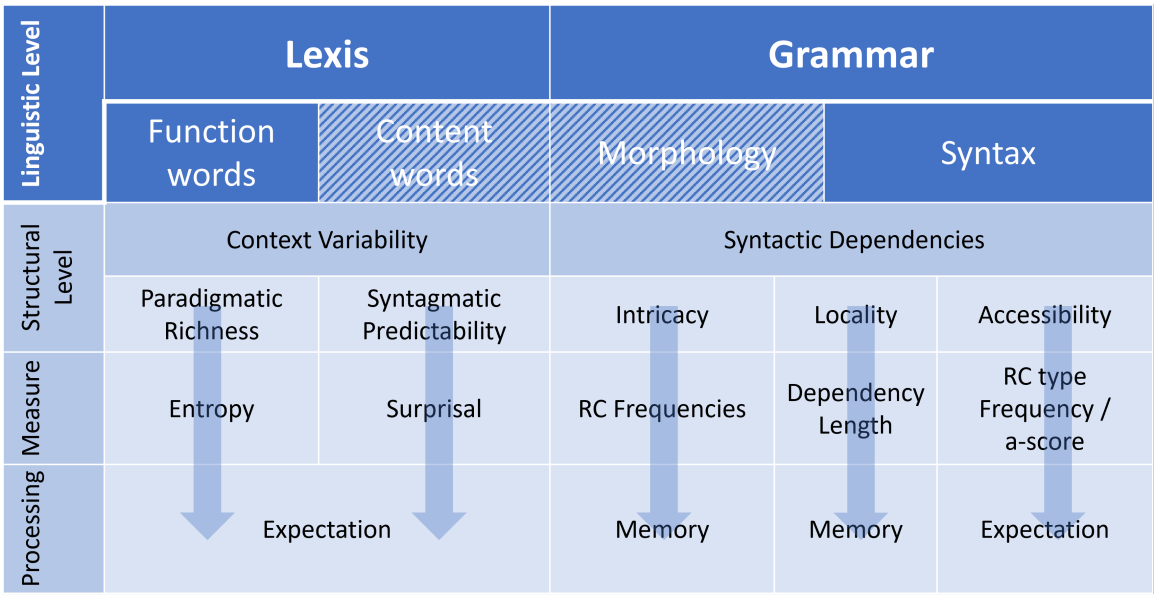


Figure 2.6: Processing-related complexity

Hence, in the present thesis, we define the following characteristics of processing-related complexity (as schematized in Figure 2.6):

1. Complexity is bound to a specific linguistic level, i.e. lexis and grammar.
2. Lexico-grammatical complexity is modulated by the variability of *syntagmatic* and *paradigmatic* contexts of words. Syntactic complexity is modulated by the composition of *syntactic dependencies* in *intricacy*, *locality*, and *accessibility*.
3. Complexity is related to processing effort arising from different cognitive constraints depending on the linguistic level. Lexico-grammatical complexity is related to *expectation-based* processing effort; syntactic complexity is related to *memory-based* as well as *expectation-based* processing effort.
4. Complexity-related processing effort can be quantified in terms of correlated measures. Processing of lexico-grammatical complexity is correlated with surprisal and entropy, and processing of syntactic complexity is correlated with dependency length and RC extraction type.

2.1.4 Efficiency, utility, complexity and diachronic change

In the preceding three sections, we have defined the concept of communicative efficiency (Section 2.1.1) and related it to the notion of *utility* (Section 2.1.2) adding

a register-theoretic aspect to the concept of efficiency. We showed that communicative efficiency is tightly connected to the concept of complexity (Section 2.1.3) and we provided the complexity definitions relevant to this thesis. In this thesis we are guided by the assumption that scientific writing has evolved towards lower (lexico-) grammatical complexity in response to lexico-semantic pressures in order to maintain communicative efficiency. Thus, in the present section, we want to embed the concepts *efficiency*, *utility*, and *complexity* in a more general theoretical framework of the diachronic evolution of scientific language. Two prominent general theories on language evolution are Haspelmath's usage-based take on Optimality Theory (OT) of diachronic adaptation (Haspelmath, 1999) and Hawkins' *Performance-Grammar Correspondence Hypothesis (PGCH)* (Hawkins, 2004). Both address the question of how languages evolve according to "preferences of performance" (Hawkins, 2004, p. 148) and "gradually become fixed conventions" (ibid.). The theories understand linguistic evolution in a Darwinian way, where "only the preferred structure is generated and dispreferred options are eliminated altogether" (ibid). According to Hawkins (2004), there is a close relationship between performance and grammars, i.e. grammatical evolution is the result of preferences of usage. He proposes three principles that involve complexity as an adjusting screw of efficiency (Hawkins, 2003, p. 121): *Minimize Domains* (Hawkins, 2003, p. 123), i.e. keeping syntactic relations as short as possible; *Minimize Forms* (Hawkins, 2003, p. 135), i.e. formal reduction, which is of advantage for processing "as long as the relevant information can be recovered, from discourse, from real-world knowledge, or from some accessible linguistic structure" (Hawkins, 2003, p. 136) and *Maximize Online Processing*, i.e. predicting the general avoidance of syntactic structures that might cause a delay in processing or the need to "look ahead" (Hawkins, 2003, p. 144). In Hawkins' approach, "grammars and grammatical evolution" are seen as complex adaptive systems (Gell-Mann, 1992) in which efficiency and processing ease are the driving forces pushing language to adapt to changes (cf. Hawkins, 2003, p. 143). According to this general theory of language evolution, grammars evolve in accordance with preferred choices in linguistic performance. These preferred structures are assumed to rely on cognitively motivated preferences for efficient language use: syntactically and lexically as short and unambiguous as possible. Although these assumptions refer to the evolution of entire languages (as opposed to sub-languages such as scientific language), the ideas are inherently functional in the way that they connect linguistic evolution to usage preferences determined by users' communicative needs. The functional character of striving for efficiency is stressed by Mufwene et al. (2017, p. 7) by pointing to the fact that languages are in a constant state of flux and evolution, as they respond to a variety of communication pressures and demands that are shaped by the speakers and the particular contexts in which they interact. As languages evolve, they undergo a continual process of change, with new features emerging and older ones evolving or vanishing. Given the dynamic nature of languages, it is reasonable to investigate how their complexity changes under various ecological pressures. By viewing languages or

their subsystems as complex dynamical systems, we can gain a deeper understanding of their behavior and evolution over time.

Having said that, language evolution is strongly situational and therefore inherently linked to register formation. For instance, the principle of *Reduce Forms* only holds given the condition that the intended meaning can be recovered from intra- or extra-linguistic context. This is perfectly in line with the principle of *utility* (Jaeger & Tily, 2011) as well as the conditions for *formal reduction* given by Levshina (2018, p. 60), both stating that efficiency is conditioned by linguistic function. So, rather than assuming a general change valid for the entire usage of a language, we assume that registers change independently of one another. In consequence, we assume that the degree to which formal reduction happens in a language is strongly dependent on the specific language, text type, and register that communication takes place in, i.e., while in literary texts aiming to entertain (Decker, 1974) stronger variation (and creativity) in expressive structures is more natural, scientific language mainly aiming to inform (Weaver & Kintsch, 1991) probably shows a stronger tendency towards the minimization of forms.

The general aim of this thesis is to show how the use of relative clauses has evolved over time, as well as how that change in use may contribute to improved efficiency in scientific discourse and possibly counterbalance the continuous pressure for lexical innovation. As stated earlier, relative clauses are elaborate constructions describing noun phrases by means of clausal post-modification. Relative clauses are therefore constructions combining lexical as well as syntactic aspects. On the lexical side, relativizers as the introductory elements can be analyzed with regard to the variability of syntagmatic and paradigmatic contexts they appear in. The degrees of lexical variability of relativizers and the resulting lexico-grammatical complexity can be assumed to affect expectation-based processing cost, where lower variability is assumed to lead to lower processing cost due to higher predictability of the forms that usage converges on. Being clausal post-modifications of noun phrases, relative clauses can also be inspected with regard to their syntactic variability. In terms of working memory, they are extremely wasteful. However, in some situations, relative clauses are necessary to transmit a message, since they provide the receiver with essential background information. Having in mind that over time the shared knowledge of a scientific community increases, and following the rationale of efficiency, it would be plausible to think that the use of RCs should decrease and only those kinds of RCs should survive that are necessary for the successful transmission of a message. Also, the extraction type of a relative clause (i.e. subject-extracted, object-extracted, etc.) can be analyzed as an indicator of processing difficulty that can be modulated in favor of successful scientific communication. The general assumption is thus that RCs are an adequate linguistic means to achieve a trade-off between lexical and syntactic complexity for successful scientific communication.

Finally, why is it interesting to conduct this analysis in two languages? The answer to this question involves two perspectives. The first pulls in the direction of univer-

sality, i.e. whether complexity reduction for efficiency improvement applies in the same way and to the same extent cross-linguistically. The other has a socio-historical background. The two language areas (the Anglo-Saxon and the German) have largely different historical courses, including the development of scientific activities and their institutionalization. In this way, it is particularly interesting to see if and how two evidently different socio-cultural developments affect linguistic developments. In the second part of this chapter, we will therefore sketch major extra-linguistic developments that may have influenced the creation and advancement of the scientific meta-register.

2.2 Extra-linguistic factors in the development of the scientific meta-register

In this thesis, we investigate the development of scientific language as a meta-register focusing on the time span from 1650 through 1899. In the present section, we would like to motivate the choice of the time span by giving a glimpse into the historical background of the period with a special view of the development of the scientific meta-register. To do this, it is important to put linguistic developments in their historical context. Looking at two languages, English and German, the historical background is certainly mostly different; however, major cultural developments (such as the *Scientific Revolution* and *Industrial Revolution*) may concern both languages to a similar extent, albeit at varying points in time. We will examine the socio-historical background of the *United Kingdom* and of the *Holy Roman Empire of the German Nation*, and compare them by looking at three major extra-linguistic factors affecting the development of the scientific meta-register, such as:

1. Institutionalization of science
2. Standardization of the vernacular
3. Scientific publishing practice in the vernaculars

2.2.1 The Scientific Revolution and the institutionalization of science

The first 50 years of the period (i.e. the second half of the 17^{th}c.) coincide with the last decades of the *Scientific Revolution* marking the beginning of modern science as we know it today. For this reason, the period is ideal for observing the formation of scientific language as a meta-register from its very beginning. The Scientific Revolution is generally described to take place between the late 16^{th} and early 18^{th} c. (Shapin, 2018) and represents a deep socio-cultural transformation in Europe affecting the way people thought about nature and the world, leading to fundamentally

new approaches toward science, and also leading to significant socio-cultural transformation in Europe. This transformation is generally assumed to have changed people's perceptions of nature and the world and brought about the emergence of entirely new scientific methodologies (ibid.) as science became an autonomous field separate from philosophy and technology (Hall, 1954). The most substantial transformations occurring in the Scientific Revolution included a turn towards experimental and empirical research, using "quantitative" rather than "qualitative" methods (Brush et al., 2019), and were led by "rationalism" instead of "emotionalism" (cf. Biber & Finegan, 1989, p. 512). During the Scientific Revolution, not only did the quality of science change dramatically, but also its quantity, leading to an exponential increase in scientific activity and an increasing amount of scientific publications. This boost in scientific activity and a need for forums of scientific discussion and exchange made it necessary to create institutions to organize scientific life. To address this need, scientific societies were established. The German *Academia Naturae Curiosum* was founded in 1652 by doctors and natural scientists in Schweinfurt and is the "oldest continuously active scientific society" (Teich, 2015). In 1677, the society was promoted from private to imperial authority and its name was changed to *Academia Caesarea Leopoldina* (Pörksen, 1986), still known today under the short name *Leopoldina*. Only a decade after the *Leopoldina*, the *Royal Society of London for Improving Natural Knowledge* was "created by royal charter in 1662" (Brush et al., 2019). In 1700, yet another German academy uniting the natural sciences and the humanities under one roof was founded with the support of Gottfried Wilhelm Leibniz: the *Sozietät [later "Akademie"] der Wissenschaften* (Pörksen, 1986, p. 62). However, "[a]ltogether, the social conditions for the development of scientific activities were not propitious in a Germany fragmented in the aftermath of the Thirty Years War" (Teich, 2015, p. 20). According to Teich (2015, p. 20), it was not the lack of scientific experts in the Germanic countries that made the German societies less successful than the *Royal Society*, but rather "the governments of the various kingdoms and principalities [who] showed little interest in them". Pörksen (1986) also mentions the lack of a nation-state as the principal reason for the German academies never reaching the same status as that of the *Royal Society*. The importance of the Royal Society, on the other hand, is mostly due to its highly influential members (Boyle, Leeuwenhoek, Newton, Franklin, Priestley, Hunter, Black, Faraday, Maxwell, Kelvin, and Galton to name but a few) as well as its "firm control of scientific communication" and "a near-total monopoly over British and American science lasting into the 1800s" (Atkinson, 1996, p. 334). Regardless of their importance, the creation of the British *Royal Society* as well as the German societies resulted from the Scientific Revolution. They were established with the firm intent to uphold the newly established values of good scientific practice. This aspiration is even reflected in the *Royal Society's* motto *'**Nullius** addictus iurare **in verba** magistri, – quo me cumque rapit tempestas, deferor hospes'*. The Latin formula essentially means *'take nobody's word for it'* and implies the resolve of the Fellows to resist the control of authority and validate all claims through experimentation and

factual evidence (The Royal Society, 2022).

2.2.2 Standardization of the vernaculars

Late Modern English (1700–1900) as well as New High German (1650–1900) – see Section 2.3 – are generally considered stages in which the grammatical configurations of the present-day languages had already been established and relatively little change happened compared to earlier historical-linguistic stages (e.g. Aarts et al., 2012, for English). However, the period is linguistically interesting regarding achievements made in terms of building a stable and universally comprehensible vernacular equipped for all kinds of situational contexts and usages. As we will show , this development was at work in both English and German alike.

English The standardization of the vernacular was "one of the most important socio-linguistic developments affecting the Modern period" (Romaine, 1998, p. 6). For English, the 17$^{\text{th}}$ c. was still a period of linguistic experimentation and non-standardized orthography and grammar due to a flexible posture towards language use typical of Early Modern English. Drawing from an abundance of varying linguistic options, writers at the time were in search of linguistic perfection (Brinton & Arnovick, 2006). The English vernacular was, however "largely uncodified [and] unsystematized" and variation in grammatical usage was even a problem amongst educated people (Baugh & Cable, 1993, p. 251). By the end of the 17$^{\text{th}}$ c., as England had become a nation-state, there was a growing interest in standardizing the vernacular. However, in contrast to other European countries, "England continued to lament the lack of an adequate dictionary, [while] Italy and France had both apparently achieved this object through the agency of academies" (Baugh & Cable, 1993, p. 259) – Italy's *Accadmia della Crusca* was founded in 1582 and published a dictionary of Italian in 1612, and the French *l'Académie francaise* was founded in 1634 and published a dictionary in 1694 (Baugh & Cable, 1993). Robert Hooke, a polymath and member of the *Royal Society*, unsuccessfully advocated the foundation of an English language academy in 1660; the *Royal Society* itself showed no interest in linguistic questions (cf. Baugh & Cable, 1993, p. 260). In the 18$^{\text{th}}$ c., then, the growing unease about the linguistic non-uniformity as well as rationalist philosophical beliefs that language, like everything else, ought to be logical, orderly, and symmetrical led to increased efforts to standardize the vernacular. English grammarians at the time aimed for a one-to-one relationship between form and meaning by avoiding redundancies and "alternative forms with the same meaning, or multiple meanings for the same form" (Brinton & Arnovick, 2006, p. 357 ff.). They were led by three general principles: "(1) to reduce the language to rule and set up a standard to correct usage; (2) to refine it – that is, to remove supposed defects and introduce certain improvements; and (3) to fix it permanently in the desired form." (Baugh & Cable, 1993, p. 252). The quest for standardizing the English vernacular thus included the conservation of efficient and proven linguistic "features of rational discourse" (Baugh & Cable, 1993, p. 245) and

led to the publication of early grammar books in the 1760s, such as *"The Rudiments of English Grammar"* published in 1761 by Joseph Priestley and *"Robert Lowth's Short Introduction to English Grammar"* (cf. Baugh & Cable, 1993, p. 269). The rise of prescriptivism of course not only concerned grammar but also triggered a reform of spelling and the extension of the vocabulary. With the increase in technological achievements and scientific discoveries, there was a constant growth in vocabulary which had to be organized and documented, leading to the publication of several dictionaries (Baugh & Cable, 1993, p. 260) and culminating in the (at the time) most comprehensive *Johnson's Dictionary* in 1755, compiled by Samuel Johnson.

German Also in the German-speaking countries, there had been a growing consciousness about the heterogeneity of the vernacular since the late 16th c.; however, both politically and linguistically, there was no unity in the German-speaking area until well into the 19th c., making the path to standardization even stonier. During the first half of the 17th c., however, four different varieties of written German ("Ostmitteldeutsche Schriftsprache" - Eastern Middle German, "Oberdeutsche Schriftsprache" - "Upper German, Schweizerdeutsche Schriftsprache" - Swiss German, "Niederländische Schriftsprache" - Dutch) (cf. Ernst, 2021, p. 190), still coexisted. Thus, the beginning of the period we are looking at here is characterized by both the absence of a generally accepted linguistic norm on the one hand, and on the other hand an increasing wish to standardize the German language to find a norm on all linguistic levels (cf. Ernst, 2021, p. 174). Factors that helped promote the general wish for standardization were the gradually changing conditions of linguistic change, such as an exponential increase in literary (print) sources due to the invention of the letterpress, as well as an increasing influence of early grammarians (such as Gottsched and Adelung) and the foundation of linguistic societies (cf. Ernst, 2021, p. 174). As a result of the above factors, during the period between 1650 and 1900, the cornerstones of present-day German were laid. The strongest pressure towards standardization of the German language came from the members of the educated aristocracy, who out of a nationalist motivation sought a unified language. Also, there was a growing influence of other European languages such as French, Italian, and Spanish. Especially French and Italian served as examples due to their early standardization through the national language academies. Through a strong exposure of German to the French language at court, a great many French words were integrated into the German lexicon. At the beginning of the 17th c., however, especially learned noble people (including Leibniz) came together with the aim of counteracting this development and pursuing two aims: the creation of a uniform written vernacular, and the creation of a vernacular free from foreign words. In an institutionalized form, these aspirations were further pursued with the creation of language societies (cf. Ernst, 2021, p. 179). The first and most influential language society founded in 1617 was the *Fruchtbringende Gesellschaft*, also known as *Palmenorden* (cf. Ernst, 2021, p. 180). One of the most important outcomes of the society and a decisive event for the linguistic historiography of the German language was the epoch-making Schot-

tel's (1663) *"Ausfuehrliche Arbeit von der Teutschen Haubtsprache"*. In contrast to modern textbooks for language teaching, however, 16th and 17th c. textbooks were still strongly influenced by Latin serving as a metalanguage, i.e. the meaning of words was described in Latin instead of German (cf. Ernst, 2021, p. 185). The two most important figures in standardizing the vernacular were Johann Christoph Gottsched (1700–1766) and Johann Christoph Adelung (1732–1806). Gottsched integrated the ideals of enlightenment philosophy in the linguistic norms he proclaimed by advocating a 'natural' language, which first and foremost had to be clear and unambiguous, i.e. syntax should not be complicated, and the lexicon had to be free of dialect words, foreign or archaic words, neologisms or metaphors (Ernst, 2021, p. 188).

2.2.3 Scientific publishing practice in the vernaculars

The beginning of our time period (mid-17$^{\text{th}}$ c.) was not only the beginning of modern science, but also the beginning of the vernacular replacing Latin as the language of scientific communication. In this regard, English scientific writing was far ahead of scientific writing from German-speaking countries. While in the English vernacular had already been well established in the 17$^{\text{th}}$ c., especially due to the Royal Society's decision to use English as its primary language of communication, the situation in German-speaking countries was much different. Here, the shift toward the German vernacular for scientific publications did not come overnight. It took 300 years. A quantitative review of publication numbers in German and Latin shows that in 1518, the proportion of book production in German was only 10%, whereas at the end of the 18$^{\text{th}}$ c. only 4–5% was in Latin (Maas, 2012, citing Bach 1965, p. 309). Several factors played a role in the transition. Early motivations came from practical necessity. For instance, the demand for technical specifications coming from the practical sector of craftsmen gave rise to the publication of instructional texts with scientific content, such as mathematical measurement instructions for painters or barrel makers. Albrecht Dürer published his 'Underweysung der Messung' in 1525 in German, including the first version of German mathematical terminology, followed by Kepler's 'Messekunst Archimedis' in 1616 (Ernst, 2021). Although publications with mathematical content were published in German early on, the real breakthrough of the vernacular in the natural sciences came only when German universities started to teach and write in German. However, in German universities, the transition from Latin to the vernacular in the natural sciences came relatively late compared to other countries such as Italy, France, and England, namely in the second half of the 18$^{\text{th}}$ c. This delay was not only confined to the university context; German was also lagging behind in introducing the vernacular in scientific communication, such as in journal articles. Due to the increasing scientific activity and an increasing scientific community as a result of the Scientific Revolution, new ways of disseminating the achieved results had to be found. With the growing number of scientifically interested people and the speed of scientific text production, scientific results could no longer be pub-

lished in expensive books (cf. Brush et al., 2019). For this reason, the established scientific societies began to publish scientific papers in their own proceedings and journals (Brush et al., 2019). Due to the close connection between institutions and their journals, the means of scientific dissemination were becoming increasingly standardized and more easily accessible to a wider public. France was the first country to publish a scientific journal in a vernacular language, the *"Journal des Scavans"*, even two months before in the UK, the Royal Society started publishing its *"Philosophical Transactions of the Royal Society"* in 1665. It was Henry Oldenburg, the first editor of *Philosophical Transactions of the Royal Society*, who "effectively invented the modern scientific journal; and for the next 200 years it was [...] the single most influential such journal in the world" (Bazerman, 1988, Chap. 5, cited after Atkinson 1996, p. 335). It is important to know, though, that during the first 100 years of their existence, the *Philosophical Transactions* were not edited by the *Royal Society* itself but by independent editors like Oldenburg, wealthy enough to carry the journal financially (Fyfe et al., 2020). In 1752, *Transactions* were taken over by the *Royal Society*. This takeover brought with it drastic changes in the publishing practice of the journal. While before, contributions to the journal had been read (if at all) by a serving Secretary of the Society, after the takeover, a "more systematic reviewing of papers" for acceptance of a contribution to the journal was anchored in the constitution of *Transactions* by the mid-nineteenth century (Fyfe et al., 2020, p. 4). Due to the increased accessibility to the wider public, English-speaking scientists began to adapt their writing by embracing a new "ideal of universal comprehensibility" requiring "new precision in language and a willingness to share experimental or observational methods" (Brush et al., 2019). For instance, the *Royal Society* itself actively influenced the development of the scientific meta-register by giving instructions for the stylistic specifications of scientific texts. In *"The History of the Royal Society"*, Thomas Sprat mentions that according to the *Royal Society's* ideal, texts should avoid "amplifications, digressions, and swellings of style". Instead, scientific texts should be characterized by a "primitive purity and shortness" and a "close, naked, natural way of speaking" (Sprat 1734, cited after Biber & Finegan, 1989, p. 512). To the present day, *Philosophical Transactions* represent continuous documentation of the English scientific meta-register.

In Germany, the development was much different, however. Although the *Leopoldina* started publishing its own journal only five years after the Royal Society (in 1670), its annual publication, the *"Miscellanea curiosa medico-physica Academiae Naturae Curiosorum sive Ephemeridum medico-physicarum Germanicarum curiosarum"* was exclusively written in Latin for a long time. Also, its Latin name was changed five times before it had its first German name *"Nova Acta Leopoldina: Abhandlungen der Deutschen Akademie der Naturforscher Leopoldina"* in 1928, and publications of the in-house journal were written in Latin until the 19th c. The transition from the purely Latin publication through the end of the 18$^{\text{th}}$ c. to a partly German publication can be observed from the volumes of 1818 onward. One reason for the journal

to stick to Latin for such a long time is the journal's initial focus on medical topics: Latin was generally regarded as the language for scientific medical communication, while only surgeons and doctors with no university academic training used the vernacular (Maclean, 1963). It is important to note that this transition is probably the most influential factor when it comes to the formation of the German scientific meta-register, since register and language are inextricably intertwined. The strong adhesion to Latin in academic circles thus stalled an institutionalized development of a German vernacular language of science. In this way, the nature of institutionalized publishing practices influenced the development of the scientific meta-register to a great extent.

2.2.4 Summary

We have learned above that society, scientific life, and with it, the conditions for scientific text production have changed dramatically throughout the years between 1650 and 1900. The Scientific Revolution led to the creation of scientific institutions, and out of these institutions, the publishing of scientific articles became institutionalized through in-house journals. While institutionalization and publishing practice have evolved mostly in parallel in German and English, the biggest divergence between the two is the actual beginning of scientific writing in the respective vernacular. German vernacular scientific discourse actually only truly started at the beginning of the 19^{th}c. – a good 100 years after English scientific writing. While not directly connected to linguistic developments, socio-economic factors such as the establishment of a nation-state and the industrial revolution (IR) of course had an impact on the development of scientific writing, too. The IR beginning in England in the 18^{th} and 19^{th} c. was the most important driver of socio-economic revolutions, leading to the foundation of polytechnic universities and an explosion in the natural sciences with the development of new sub-disciplines and vast scientific text production (Drozd & Seibicke, 1973). The growth in prosperity arising from the IR led to "a growth in the middle class and an increase in social mobility" (Brinton & Arnovick, 2006, p. 357) as well as "the rise of a popular, middle-class literacy" (Biber, 1995). de Courson & Baumard (2019) find prosperity seems to be the key factor for scientific progress (measured in the number of scientific publications) per country. The fact that the IR started in England and only arrived in Germany almost 100 years later can be assumed to have contributed to the general development of scientific vernacular writing in German-speaking countries. We can thus assume that both socio-economic factors and the delayed institutionalization of the vernacular in German academia and scientific publishing are key factors for the time-shifted beginning of German vernacular scientific writing and hence a later development of a German scientific meta-register. In Section 2.3, we will thus have a specific look at the development of the vernacular languages between 1650 and 1900 and specifically of the scientific meta-register.

2.3 Linguistic change and the formation of the scientific meta-register

In this Section, we will first look at general linguistic changes in German and English in the period between 1650 and 1900 (Section 2.3.1). We will look at both languages separately, starting with English. We will then specifically focus on existing research on the development of the scientific meta-register during this period of time (Section 2.3.2) and in Section 2.3.3, we will review existing research on relative clauses (RCs) in scientific writing.

2.3.1 General linguistic change between 1650 and 1900

2.3.1.1 English

In terms of linguistic periodization, strictly speaking, the time period we are looking at in this thesis spans two language stages: the very last decades of Early Modern English (eModE), generally referred to by historians as spanning the period 1500–1700 (Romaine, 1998, p. 6) and the entire period of Late Modern English (lModE, 1700–1900). But what were the actual changes in the English language? In terms of syntax, global changes distinguishing Old English (OE) from Present-Day English (PDE) had already been undergone by 1776, leading to relatively fixed word order (SVO/SVcomplement) (Romaine, 1998, p. 170 f.). This, however, does not mean that the English language did not change at all in the lModE period, but the syntactic change was rather statistical in nature, with a given construction occurring throughout the period and either becoming more or less common generally or in particular registers Aarts et al. (2012, and colleagues). The latter provides a non-exhaustive overview of the major syntactic changes in the lModE period, stating that lModE was an interim stage between eModE and PDE. The authors distinguish between "categorical and statistical changes". Apart from the final steps in the consolidation of the SVO order, categorical changes in lModE include the domain of voice, with the introduction of the progressive passive (e.g. *was being introduced*) and the 'get-passive'. Amongst the most notable statistical changes, Aarts et al. (2012, p. 870) name the "consolidation of the progressive, the decline of the *be*-perfect and the regulation of periphrastic *do*" as well as a trend towards avoiding preposition stranding. Another important statistical shift is the "Great Complement Shift" from OE to PDE (Rohdenburg, 2006, p. 143) describing the replacement of finite complementation strategies by non-finite ones, mostly by infinitives. Within non-finite complementation strategies, however, the trend shifted from infinitival complementation to *-ing* complements (Aarts et al., 2012). Rohdenburg (1996, 2006) also looked at the conditions of the retention of finite instead of non-finite complements and gave a processing-related explanation for this development. He found that *that*-complement clauses are "easier to process than non-finite structures" and are therefore preferred in contexts that are cognitively complex,

i.e. contexts containing intervening material or negations (cf. Rohdenburg, 2006). Regarding complementation, Aarts et al. (2012) also mention the consolidation of the relativizer paradigm. During the lModE period, relativizers, which before had been used in relatively arbitrary ways, became consolidated in particular contexts of usage. For instance, *which* had formerly been used for human as well as non-human referents, but became restricted to non-human referents in lModE. For a more detailed description see Section 2.3.3. Now that all systemic changes had been completed at the earlier language stages, change in lModE instead tends to affect specific patterns of choice, which contribute to making the English vernacular more efficient by finding shorter ways of expression when cognitively possible.

2.3.1.2 German

In terms of German periodization, the time between 1650 and 1900 is commonly referred to as New High German (NHG). However, as in most historical linguistic definitions of language stages, there are disputes about the exact beginning and end of the stage (for an overview see Hartweg & Wegera, 2005, p. 21). Konopka, for instance, sets the beginning of NHG at the beginning of the 18$^{\text{th}}$ c. and the end around the 1830s (Konopka, 1996, p. 15). However, we are following the (most widely acknowledged) periodization of Scherer (1875) and Eggers (1977, 1986), who set the beginning of NHG in the mid 17$^{\text{th}}$ c. In many attempts at periodization, NHG is subdivided into Early New High German (ENHG) and NHG, the first representing an interim period between Middle High German (MHG) and NHG. For Eggers and Scherer, the interim period ENHG ends and NHG begins with the publication of Schottel's Grammar ('Teutsche Sprachkunst', Schottel, 1967) in 1641[3] (Hartweg & Wegera, 2005, p. 23). Eggers and Scherer do not set an end to the NHG period, assuming that the period continues up to the present day (or in Scherer's case, his lifetime). Other periodizations do set a beginning for Present-Day German (PDG), albeit at differing points in time. For instance, Bräuer (2001) sets a limit after 1950, calling the following period "Gegenwartsdeutsch"; von Polenz (1978) differentiates between "älteres Neuhochdeutsch" (older NHG) and "Deutsch im 19. und 20. Jh." (19$^{\text{th}}$/20$^{\text{th}}$ century German) beginning around 1850, and for Keller (1978), "The Modern Standard German" begins as early as 1800. For the present thesis, we will use the periodization by Eggers and Scherer and call our period of observation NHD until 1900.

According to Admoni (1985, p. 1545), in the 18$^{\text{th}}$ c. the grammatical system of the German language was fixed in much the same form as it exists today. He states that the most important variants in the quantitative sentence design were also determined in this period. The changes we can observe in this time period are thus of a statistical rather than categorical nature. (Semenjuk, 1972, p. 141 ff.) reports a clear tendency towards a reduction of grammatical variation in the sense of a stronger trend towards

[3] or the publication of the revised edition in 1667.

normalization in the 18th c., possibly due to the territorial unification ("territoriale Vereinheitlichung im 18. Jahrh.") (Piirainen, 1980, p. 598). Regarding syntactic developments, Admoni (1972, p. 18) reports that the total length of sentences is highest in the 17th c., especially in legal and specialized language, as a result of increasing differentiation between written and spoken discourse. The extreme sentence lengths achieved in this period are due to an increase in hypo- and parataxis. Both the number and the length of the subclauses within one sentence affect the overall sentence length (Admoni, 1985, p. 1541). Especially the increase in hypotactic structures is the result of a strong Latin influence on specialized discourse (Scaglione, 1981), while the flourishing of hypotaxis in the 17th c. can be interpreted as a trait of baroque style (Konopka, 1996, p. 21). The overloaded sentences resulting from this massive use of clause complexes can be assumed to have had negative consequences on their processing. For instance, Betten (1987, p. 74) notes that hypotaxis is cognitively highly demanding and might be especially hard to understand for non-academic readers. The inconvenience of the hypotactic style is also mentioned by Admoni (1985, p. 1540), who observes an urge to use hypotaxis resulting in ungrammatical constructions, which sometimes come without a main clause and appear as a juxtaposition of interdependent subclauses. In the 18th c., sentence lengths decrease notably due to a strong decrease in hypotaxis (Admoni, 1972, p. 297). Only in scientific literature do the sentences stay comparatively long (Admoni, 1985, p. 1543); shorter sentences are found in fictional texts. The turn away from the extreme sentence lengths can be interpreted as a response to the ideals of enlightenment (Scaglione, 1981) such as naturalness and comprehensibility (Betten, 1987, p. 75). However, long sentences do not seem to be abolished entirely. As a result, at the end of the 18th c., two different synchronic paradigms with both the simple and the intricate style had been established (cf. Schildt, 1984, p. 172), (Tschirch, 1989, p. 225), (Wolff, 1991, p. 146) and still seem to exist today (Schildt, 1984, p. 196, 214). The ongoing reduction of sentence length over time can, however, be regarded as a general trend beginning in the 18th c. and continuing until the present day (Schildt, 1984, p. 250). At the same time, a complementary development can be observed: While in the 18th c. sentences start to become shorter, clauses become longer (Admoni, 1985, p. 1544), leading to overly long clauses with many nominal groups in the 19th and 20th c. (Admoni, 1985, p. 1545f.). In the 19th and 20th c., clauses stay stable in length, but the length and complexity of the sentence decrease, especially in scientific writing, while in fictional texts clauses become shorter (Admoni, 1985, p. 1549f.). The extension of the clause is due to the extension of the nominal group, which according to Admoni (1972) is only possible due to the precision of the morphological system in the 17th and 18th c. Functionally, the densification of the nominal group can be regarded as a result of the urge for information density and conciseness (Erben, 1984, p. 180).

2.3.2 The evolution of the scientific meta-register

Apart from the general trend towards the standardization of the vernaculars, the newly established way of doing science brought about by the Scientific Revolution also led to substantial changes in the scientific meta-register over the past 300 years (Biber & Finegan, 1989; Atkinson, 1998; Biber et al., 2009; Biber & Gray, 2016). For instance, in the 16th and 17th c., scientific texts were mostly combinations of knowledge deriving from craftsmen's practical, even experimental, experience and classical, i.e. rather theoretical methods. This combination of the practical technical-artistic and classical scientific approaches subsequently led to a shift in scientific text production, becoming manifest in terms of the texts' composition, presentation, argumentation, and terminology, as well as the combination of text and sketches; see Drozd & Seibicke (1973, p. 17) and Pörksen (1986, p. 12)). In the late 17th c., scientific articles were often written in the form of "letters to the editors of a journal" and represented highly explicit accounts of the actions taken by the author and the observations made during the process (Biber & Gray, 2016, p. 51). Also, most accounts were written by researchers of no particular specialization, and their target groups were mostly other generalist researchers, even including the broader literate audience (Biber & Gray, 2016, p. 51). Over time, a diversification of the scientific meta-register started to emerge. It was, however, not before the 18th c. that natural-scientific sub-disciplines such as chemistry and physics were born, chemical, zoological and botanical nomenclatures were created, and natural-scientific principles determined the methods (Drozd & Seibicke, 1973, p. 17). Important international protagonists of the new specialized scientific sublanguages were Linné, Newton, and Lavoisier (Pörksen, 1986, p. 22). At the beginning, the founders of the new disciplines wrote in a relatively general style, comprehensible by a broad public, which was due to the fact that a specialized community simply did not yet exist at the time. They had to communicate their new discoveries by establishing new concepts with their respective terminology, without assuming any of it to be known to their readers. This required a language that introduces and explains new concepts, taking up old knowledge and connecting it to new concepts to further develop new theories. In addition to general linguistic standardization processes (e.g. in orthography and grammar), in the 18th c., the scientific article itself became standardized in terms of its organizational structure. Due to a shift in the scientific research design during the 19th c. (i.e. from observational to experimental), the organizational structure of the scientific article also became conventionalized in the modern format 'Introduction, Methods, Results, and Discussion' (cf. Atkinson, 1998, p. 70).

2.3.2.1 English

Being the most prolific and extensive body of English scientific writing covering the time between the late 17th onward, the *"Philosophical Transactions, and Proceedings of the Royal Society"* (PTPRS) are a continuous source of scientific writing. For this

reason, much of the previous work on scientific English in the lModE period is based on the texts published in the PTPRS (e.g. Atkinson, 1992, 1998, 1996; Dear, 1985; Bazerman, 1988; Valle, 1997; Degaetano-Ortlieb & Teich, 2016, 2018; Teich et al., 2021). Early studies (e.g. Bazerman, 1988; Atkinson, 1992) are rather qualitative in nature and report an increasing orientation towards experimental reports as a central component of scientific communication (starting with a low proportion of as little as 5% of the total publication load). Such reports over time also changed in style from pure descriptions, becoming more methodological and finally focusing on theory-testing and evaluation by experiment. Also, Valle (1997) reports that texts evolve from being rather descriptive and narrative in earlier periods and over time becoming more "systematic" and "argumentative" (cited after Atkinson, 1996, p. 336). More quantitative (frequency-based) register analyses in (Atkinson, 1992, 1996) report an evolution of the scientific meta-register along Biber (1988)'s dimensions of register analysis. Atkinson (1996) finds a clear trend on three dimensions: Scientific texts develop towards an increasingly informational, non-narrative and abstract style. Diachronic studies inspecting the scientific meta-register have also been done by comparing the development of different registers to each other (e.g. Biber & Finegan, 1989, 1997; Biber & Clark, 2002; Hundt et al., 2012). Most of these studies are based on the ARCHER corpus (A Representative Corpus of Historical English Registers, Yáñez-Bouza, 2011). The corpus is a diverse compilation of historical texts from various genres, amounting to around 1.8 million words in British and American English. The timeframe covered in this corpus is from 1650 to 1999. These studies unanimously report a trend from a formerly more elaborated syntax that is more verbal and grammatically explicit (Biber et al., 2009), towards a more nominal style with stronger syntactic compression especially within the noun phrase (Halliday, 1988; Biber & Clark, 2002; Hundt et al., 2012). This compression is achieved through compressed nominal modifiers such as premodifiers and phrasal postmodifiers (Biber & Clark, 2002; Gotti, 2003; Biber & Gray, 2011b) and was found cross-regionally for British English (BE) and American English (AE) alike by Hundt et al. (2012).

The shift from verbal (verb-based trigrams) to more nominal usage patterns (noun-based trigrams) in scientific English has also been confirmed by studies using information-theoretic measures such as surprisal and entropy, e.g. Degaetano-Ortlieb et al. (2019, p. 277). They find a higher informational load of the noun-based patterns indicating that in scientific English over time "less informative usages" (such as the complementizer *that* after mental and verbal verbs) become redundant and thus fall out of usage while more informative patterns are preserved in ***specialized*** expert-to-expert communication. Along the lines of register theory (Halliday & Ruqaiya Hasan, 1985), Degaetano-Ortlieb et al. (2019) associate this development with a shift from a "reporting to expository" discourse type, assuming that the trend is the result of increasing ***specialization*** of the scientific meta-register adapted to expert-to-expert communication. Inspecting gerunds and passives with average surprisal and Kullback-Leibler Divergence (KLD, Kullback & Leibler, 1951), Degaetano-Ortlieb & Teich (2016) find

that scientific English becomes increasingly ***conventionalized*** and less productive over time in terms of specific part-of-speech trigrams both becoming less variable in their number of types and in terms of their preceding contexts. Bizzoni et al. (2020) use pointwise KLD to detect phases of innovation and consolidation at both the lexical and the grammatical level. They find recurring peaks of lexical innovation due to new scientific discoveries (e.g. the oxygen theory in the 18th c.) as well as a trend toward "consolidation in grammatical usage". The development of grammatical constructions such as passive voice and relational verb patterns and nominal patterns with prepositions, gerunds, and relative clauses point to the development of "a uniform scientific style" (Bizzoni et al., 2020, p. 7). Teich et al. (2021) investigate the interplay between *conventionalization*, i.e. the "convergence in linguistic usage over time" and *diversification* describing the process of "linguistic items acquiring different, more specific usages/meanings" in scientific English over time. They assume that the modulation of linguistic variability is in the interest of rational communication. Using diachronic word embeddings, they find an overall decrease in paradigmatic variability as shown by overall increasing distances in a vector space between words and overall decreasing entropy driving both conventionalization of fewer linguistic options and the diversification of these options. In parallel to the general downward trend of entropy, they also report on a temporary rise in entropy of specific terminological fields during the 18th and 19th centuries (Teich et al., 2021, p. 13) pointing to waves of terminological innovation. Taking a syntactic perspective on complexity in scientific writing, Biber & Gray (2016) distinguish between two kinds of complexity: phrasal complexity and clausal complexity (cf. 2.1.3.2), the latter of which goes down in scientific writing over time. Due to the absence of verbs and other grammatical signals, higher phrasal complexity leads to lower explicitness. Such compressed structures can be extremely efficient for expert readers (Biber & Gray, 2016, p. 249). The described development furthermore seems to be dependent on the specific sub-discipline. While the above-mentioned features have become increasingly distinctive for the natural-scientific literature, scientific articles in the humanities do not seem to have changed much since the 18th c. (Biber & Gray, 2016, p. 250).

In summary, previous work has shown that the English scientific meta-register has gone through a transformation from a rather narrative (linguistically explicit) reporting to a purely expositional (highly implicit) goal orientation. At the syntactic level, this transformation can (among other features) be observed in the shift from clausal complexity towards phrasal complexity, leading to lower grammatical explicitness as a response to the increasing need for efficiency in scientific communication. This need for efficiency is part of the interplay between lexical innovation in particular scientific fields, which has been detected in terms of temporary peaks of high entropy/surprisal. To serve this increased communicative need for efficiency, scientific English has undergone a trend toward *specialization*. On the syntagmatic level, the meta-register has become more efficient by packaging information more densely as indicated by a higher informational load of nominal patterns in their syntagmatic contexts. On the

paradigmatic level, there has been a trend toward *conventionalization* as indicated by the convergence on particular paradigmatic options and *diversification* by resettling options to different contexts of use. This convergence as well as the resettlement of options manifests itself in a general decrease in entropy. Altogether, this general effort toward creating efficiency in scientific writing seems to be a register-specific trend making communication among scientific experts as efficient as possible.

2.3.2.2 German

When reviewing the existing literature on the diachronic development of scientific German, the most prominent topics described there are the creation and development of scientific terminology and the transition from Latin to the vernacular language. In contrast to studies of the English scientific language, German studies are rather descriptive and qualitative in nature, concentrating on the socio-historical factors influencing the transition from Latin to German, as well as on the protagonists promoting this transition. Pörksen (1986) summarizes the factors driving the development of the language of the sciences as follows:

1. The organization of university teaching of natural sciences leads to the preservation and flourishing of knowledge.

2. The interest of technical professionals, as well as the economic interest in scientific development in industry and technology, lead to the support of scientific academies and societies.

3. Scientific discourse is characterized by an international orientation with Latin as a Europe-wide scientific lingua franca, which is later preserved in the form of extensive adoption of Latin or Greek terms.

4. The increase in written records contributes to the institutionalization of scientific ideals.

5. The fundamental principle underlying the evolution of the natural sciences is the wave-like discovery of new knowledge and conceptual interrelations. Scientific discourse serves the communication *about*, as well as the correction *of*, the known and the reciprocal understanding of the unknown. This leads to the continuous creation of new terminology, the negotiation of old terminology, and the reinterpretation, replacement, and extension of the existing terminology. Scientific language is thus characterized by a dynamic disappearance of obsolete words on the one hand, and the creation of new vocabulary on the other. The explosion in new discoveries since the 18^{th} c. has led to a similar explosion in new terminology at an unprecedented speed.

6. The creation of terminology is an increasingly conscious activity in the 19^{th} c., taking place explicitly by means of the designation of a term to a concept ('We shall call this XY').

7. A differentiation between scientific and general language is only partially possible, namely on the level of terminology. Scientific specialized languages are subsystems of general language.

While we agree with most of the above points, we beg to differ on the very last point. While terminology is certainly a prominent distinctive feature differentiating scientific language from general language, we are convinced that there are also major developments at the grammatical level. Some of these developments are addressed in Habermann (2011)'s comprehensive account on the development of German syntax in the natural sciences between the 15th and 19th c. Her study focuses on the influence of Latin on the emerging vernacular German as a language of scientific communication. Until the early 19th c., 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 instead of parataxis. In fact, it was technical literature ("Sachprosa") of the 17th c. that first established the verb-final word order in subordinate clauses, making it possible to distinguish between main and sub-clauses (cf. Nerius, 1967). As the formation of long and embedded sentences to present complex thoughts in one sentence becomes possible, an extreme increase in hypotactic structures in the 17th and 18th c. can be observed (cf. Möslein, 1974). As a result, the scientific German (especially legal German) of the early 19th c. has the reputation of being the epitome of intricate syntax (Möslein, 1974). Starting in the first half of the 19th c., a trend of disentanglement and reduction in sentence length (Sladen, 1917), as well as a remarkable reduction in subordinate clauses and an increase in nominalizations, is described as taking place in scientific German (Möslein, 1974; Beneš, 1981). The reduction in sentence length was especially notable in natural-scientific texts as compared to texts from the humanities (Sladen, 1917). Societal developments of the time, such as the increasing influence of mass media and other European languages of science, are reflected in a new trend towards lower syntactic intricacy. The new trend towards condensation is accompanied by an extension of clause simplexes (simple clauses without a sub-clause) by means of attributes and appositions (Beneš, 1981, p. 200) as well as prepositional phrases. On a scale of compression, Beneš (1973) situates prepositional phrases between a subclause (on the less compressed end) and relational adverbs (on the most compressed end) being extremely frequent in PDG (e.g. 'Differenzialdiagnostisch besteht die Notwendigkeit einer Abgrenzung von den Masern': Beneš, 1981, p. 202). Furthermore, Beneš (1973) mentions attribution in the form of attributive sub-clauses, infinitive and participial attributes, pre- and postponed attributes, and appositions as a frequent means of condensation. As in English, the increase of nominal groups instead of sub-clauses can be assumed to be motivated by an increasing need for exactness and effort reduction in terms of dependency length (cf. Section 2.1.3.2). Also, the increasing use of compounds as an alternative to prepositional phrases (*Extraktionsverfahren* vs. *Verfahren zur Extraktion*), and nominalizations instead of sub-clauses (*Wirtschaftlichkeitsberechnung*

vs. *die Berechnung, die wir zur Ermittlung der Wirtschaftlichkeit nutzen*) results in a loss of grammatical explicitness connected to tense, mode, and number (Möslein, 1974). At the same time, the loss of expliciteness can be assumed to contribute to lower processing effort triggered by syntactic complexity. Presumably, the increasingly condensed scientific style is motivated by the aim for clarity and efficiency of expression in response to evolving communicative needs (Möslein, 1974).

For PDG language for special purposes (LSP), Roelcke (2020, p. 78) in line with Pörksen (1986) states specifically that the particular features of LSP are first and foremost to be found on the level of the vocabulary, while specific features on the level of grammar are covered to a much weaker extent. As in English, differences in grammatical usage between general and specialized texts are of quantitative nature. In terms of syntax, declarative sentences are the most frequent sentence type in LSP (Roelcke, 2020, p. 78); (Beneš, 1981, p. 191), while relative clauses are also relatively frequent (Roelcke, 2020, p. 86) to specify and determine concepts. Present-day LSP is also characterized by frequent use of attributive noun phrase modifications, such as attributive adjectives (*"das sparsame Auto"*), participial attributes (*"das Benzin sparende Auto"*), prepositional attributes (*"das Auto aus Aluminium"*), and attributive Genitives (*"der Verbrauch moderner Kleinkraftwagen, Goethes Werk"*) (Roelcke, 2020, p. 87). These constructions contribute to syntactic complexity in two ways: On the one hand, they affect the number and connection of clauses, since the clause complexity is higher in LSP than in general German and subsequently leads to a higher total sentence length quite typical of German LSP (Roelcke, 2020, p. 88), and on the other hand, this type of attributive noun phrase modification contributes to the increase in syntactic complexity on the clausal level by increasing the complexity of individual phrases.

2.3.3 Relative clauses as markers of syntactic complexity

As mentioned in Section 2.3.1, in the lModE period, "The Great Complement Shift" was at its zenith, favoring non-finite complements over finite ones (such as RCs). At the same time, the standardization of relativizer usage was pushed forward by prescriptivist movements. The lModE period was mostly a period of statistical change, i.e., some constructions became more or less frequent and started to settle into specific contexts of usage. In the present thesis, we are interested in the diachronic development of syntactic and lexico-grammatical complexity in the scientific meta-register. Since RCs represent a good 'micro-ecosystem' to observe both types of complexity, we now focus on previous work dealing with the diachronic development of relative clause use in the scientific meta-register compared to other registers in the lModE period.

2.3.3.1 English

The motivation to look at RCs and their frequencies in the scientific meta-register is that they represent one of the "different strategies in the packaging of information (phrasal vs. clausal)" (Hundt et al., 2012, p. 214), representing an "explicit, elaborated identification of referents in a text" (Biber & Finegan, 1989). RCs and their alternative noun phrase modification strategies play an important role in syntactic complexity (Biber & Clark, 2002) and they are typical of an elaborated rather than a situation-dependent reference typical of formal written prose (Biber, 1988; Biber & Finegan, 1989). Several studies have looked at RC frequencies in different registers in English as part of noun phrase (NP) complexity development (Halliday, 1988; Biber & Clark, 2002; Biber & Gray, 2011b, 2016; Biber et al., 1999; Hundt et al., 2012). Biber & Gray (2016) on sub-register distribution (textbooks, classroom teaching, conversation, and research articles) of RCs find that RCs are especially frequent in textbooks and classroom teaching (oral-like registers), while being relatively infrequent in research articles Biber & Gray (*formal written* 2016, p. 102). Across registers (fiction, newspaper, and academic writing), they find that RCs are most frequent in newspaper articles and much less so in academic writing as well as fiction, with RCs being the frequent type amongst all types of sub-clauses (Biber & Gray, 2016, p. 106). In terms of different scientific disciplines (humanities, social science, popular science, specialist science), RCs are relatively infrequent in specialist science compared to humanities writing (Biber & Gray, 2016, p. 115, p. 123). Diachronically (1725–2005), RCs overall decrease across various registers (including fiction, news, and science); however, the strongest decrease is found for scientific writing, especially between 1930 and 1960 (Biber & Gray, 2016, p. 150). In terms of sub-registers, over time, RCs have declined most in specialist natural science articles, as compared to specialist social science articles and multi-disciplinary science articles showing a lesser decline. The least palpable decline was found for historical research writing, which was still making heavy use of RCs. "As a result, relative clauses are at present considerably more common in humanities academic prose than in any of the science registers" Biber & Gray (2016, p. 162). The decline of RCs in scientific writing was also found cross-regionally comparing BE and AE (Hundt et al., 2012). The results indicate a general shift towards a more compressed style in scientific writing. Furthermore, Hundt et al. (2012) compare the development of full RCs to reduced RCs (adnominal RCs) and other strategies of NP modification, such as NP pre-modification patterns (e.g. attributive adjectives, compounds) and prepositional phrases as NP postmodifications. On the one hand, they find "a trade-off between relative clauses (decrease) and postmodifying participle clauses (increase)" (Hundt et al., 2012, p. 236); however, the other alternative NP modification strategies do not seem to have a strong effect on the overall trend "from more expanded to less expanded" syntax (Hundt et al., 2012, p. 236).

2.3.3.2 German

Also in studies on German, RCs have been identified as promoters of syntactic intricacy (Möslein, 1974; Admoni, 1990). Based on the general assumption that relative clauses are on the lower end of the condensation cline (Biber & Gray, 2016), German linguists have also looked at LSP and scientific language with regard to the distribution of RCs and other forms of NP modification. For instance, Beneš (1973) claims that the degree of condensation increases depending on whether a matter is expressed by means of a main or a sub-clause, by an infinitive or participial construction functioning as a sentence, or by a phrase or a part of a phrase. The syntactic construction is denser and tighter, the less independent a sentence component is in relation to the predicate. In PDG technical style ('*Fachstil*'), generally, a more condensed means of expression is preferred (Beneš, 1973, p. 40 f.). Beneš provides a list of constructions representing a condensed style, amongst which, however, he also mentions RCs in the second position, stating that RCs are a common syntactic feature in LSP. On a more diachronic note[4], he claims that the trend towards stronger condensation has become visible in the recent past, reflecting an effort towards economy, conciseness, and clarity (Beneš, 1973, p. 46). Beneš also looks at the distribution of different types of sub-clauses and finds that RCs in PDG scientific texts are the most frequent, constituting 50% of all sub-clauses in their corpus of scientific texts from the second half of the 20th c. Beneš (1981, p. 190). We are, however, not aware of any studies looking at the development of RC use throughout the NHG period in German scientific texts. We therefore hope to close this gap with the present thesis targeting the diachronic development of RC use *in* and its contribution *to* syntactic complexity in scientific writing.

2.3.4 Relativizers as a markers of lexico-grammatical complexity

Having reviewed extant work on RCs as markers of grammatical complexity and the development of their usage in scientific writing over time, we will now zoom in on the diachronic development of relativizer usage in English and German scientific writing as they represent markers of lexico-grammatical complexity.

2.3.4.1 English

Before reviewing existing work describing the diachronic development of the introductory markers of RCs, i.e. relativizers in the lModE period, let us define what relativizers actually are. Formally, relativizers do not represent a word class on their own. Relativizers (or relative pronouns) are defined by Biber et al. (1999, p. 71)

[4]comparing encyclopedia entries from 1935 and 1957.

as pronouns "which mark identity with a preceding noun phrase". In Standard English, they count eight different relativizers: "*which*, *who*, *whom*, *whose*, *that*, *where*, *when*, and *why*", the latter three of which they call "relative adverbs" (Biber et al., 1999, p. 608). The group of PDE Relativizers is mostly confined to *that*, *which*, and *who*(*m/se*). However, the choice amongst these three main relativizers is most significantly influenced by the register, restrictiveness of the RC, animacy of the head noun, and to some extent regional varieties (BE vs. AE) (Biber et al., 1999, p. 616). Regarding register, *which* is associated with "conservative, academic" styles and "thus preferred in academic prose" (Biber et al., 1999, p. 616). Specifically, since the focus in the present thesis is on the two most frequent relativizers in PDE, i.e. *that* and *which* as well as a group of *wh*-relativizers belonging to the class of pronominal adverbs (e.g. *whereby*, *whereof*, etc.), we will primarily focus here on these three types of relativizers. Relative clauses (RCs), and especially the choice of relativizers, have long been a widely studied topic in English diachronic studies, covering all periods of the English language, such as the OE and ME period (Suárez-Gómez, 2012, 2008), the eModE period (Rydén, 1966; Dekeyser, 1984; Rissanen, 1984; Nevalainen & Raumolin-Brunberg, 2002) and LModE (Hundt et al., 2012; Johansson, 2006, 2012; Huber, 2017). Relativizer choice has also been studied in vernacular varieties of English (Romaine, 1980, 1982; Tottie, 1995; Tottie & Harvie, 2000; Tagliamonte, 2002; Tagliamonte et al., 2005; Levey, 2006) and in spoken and written modes (Guy & Bayley, 1995). To understand why the choice of relativizers is diachronically interesting, it is important to know that PDE relativizers (*which*, *that*, *who*(*m/se*) and *zero*) have different historical origins. While *that* and *zero* have long existed, dating back to OE, the *wh*-relativizers joined the group much later. *Wh*-relativizers were formerly used as interrogative pronouns and, inspired by the romance languages, came to be integrated as relative pronouns in Early Middle English (Mustanoja, 1960; Romaine, 1980). Until the beginning of the EmodE period, *which* had no clear-cut (animacy) differentiation between personal and non-personal antecedents (Dekeyser, 1984; Ball, 1996; Görlach, 2001; Johansson, 2012, cited by Huber 2017, p. 76). The limitation of *which* to be exclusively used with non-personal antecedents was established by 1700 (Dekeyser, 1984, p. 71). Being a "foreign" relativizer, at earlier stages, *wh*-relativizers were primarily used in complex formal, written English and only gradually came to be used in less formal language as well (see e.g. Dekeyser, 1984; Nevalainen & Raumolin-Brunberg, 2002; Romaine, 1980). By the beginning of the lModE period, the time period of our interest, the relativizer set (*which*, *that*, *who*(*m/se* and *zero*) had been established and no changes in terms of the relativizer inventory have occurred since then (Romaine, 1982, p. 69, 71).[5] However, in terms of function and distribution, there have been quite a few changes in the past 300 years (Huber, 2017, p. 76), i.e. there are statistical changes accounted for in the lModE period in various studies

[5]This claim may be correct as long as only the relativizers *which*, *that* and *who(se/m)* are concerned. However, when pronominal adverbs are included in the group of relativizers, we will show that the inventory has dramatically changed over time.

(e.g. Dekeyser, 1986) investigating the different factors playing a role in the relativizer choice over time. These factors can be of intralinguistic (animacy of the antecedent, syntactic role) or extralinguistic (i.e. prescriptive rules, genre, mode, gender, social class of the speaker/writer) nature. For instance, Ball (1996) explores restrictive subject RCs from the 16th c. to the 20th c. looking at both intra- and extra-linguistic factors (animacy of the antecedent, syntactic role, mode, and genre). Although it is generally assumed that since the 16th c. there has been very little change in the relativizer system (cf. Romaine, 1982, p. 71), Ball (1996, p. 252) reports on a shift in the use of *that* with personal antecedents, towards the primary use with non-personal antecedents, and *who* becoming the standard relativizer for personal antecedents in lModE. Furthermore, across different genres, she finds a relative increase in the use of *which* (as opposed to *that* and *zero*) until the end of the 19th c., and a decrease thereof afterwards. The opposite development was found for *that*, which decreased during the lModE period, and in the 20th c. increased again in written texts (Leech et al., 2009, p. 227). Johansson (2006) investigates the use of relativizers in the CONCE corpus (A Corpus of Nineteenth-Century English) covering three genres: science, trials, and letters from the 19th c. considering various factors (animacy of the antecedent, syntactic role, genre, gender and social class of the speaker or writer). In scientific writing, she finds that *wh*-relativizers are more frequent than *that* in the 19th c. (Johansson, 2006, p. 137). She explains the rise and dominance of *which* over *that* with its versatility with regard to restrictiveness as well as clausal functions. *That*, in contrast, is confined to restrictive RCs only, and mostly occurs in subject RCs. Johansson (2006, p. 136) furthermore attributes the shift towards the use of *wh*-relativizers with personal antecedents to the need for "clarity of expression and conciseness required of a scientific text", which can be expressed best with the grammatical explicitness of *wh*-relativizers (signaling animacy and case: Strang, 1970; Quirk et al., 1985). However, in informal and spoken English, *wh*-relativizers have never become the dominant choice, and '[t]he spread of *who* and *which*, and the recession of *that*, are especially characteristic of a formal style of writing" (Barber, 1997, p. 213). Hundt et al. (2012) also compare the diachronic development of the relativizer choice in late modern American and British English distinguishing between restrictive and non-restrictive RCs. Their findings show that *which* is the dominant relativizer in both BE and AE scientific texts. For BE they find a steady increase in *which* over time, reflecting the BE prescriptive norm to use *which* and to avoid *that* in written formal contexts.

While the above-cited studies have shown that there seems to have been a conventionalization of the relativizer *which* in scientific texts due to its higher flexibility in clausal contexts, another factor for the preference of *which* might also be the fact that (as suggested by Aarts et al., 2012) there was a prescriptive shift away from preposition stranding (see Example (11-b)) and RCs with adverbial gaps can only be built with *which* using pied piping (see Example (11-a))

(11) a. *The house* ***in which*** *we live*

b. *The house **that** we live **in***
c. * *The house **wherein** we live*
d. * *The house **in that** we live*

The version illustrated by Example (11-a) not only became preferred over the version illustrated by Example (11-b) but also seems to have replaced the formerly frequently used pronominal adverbs (cf. Example (11-c)), which we will focus on next.

While traditionally not listed amongst the core set of English relativizers, pronominal adverbs (PAs) at earlier stages (and sometimes still today) are often used to perform the function of a relativizer, representing a synthetic variant to the analytic combination of *preposition + which* (see Examples (12) and (13)).

(12) a. *He concludes, that he hath contrived **a New Instrument, whereby** every one may give himself a Clyster without any Denudation of the parts [...] (NA, An Account of Some Books, 1668)*
b. *He concludes, that he hath contrived **a New Instrument, by means of which** every one may give himself a Clyster without any Denudation of the parts [...]* (generated alternative)

(13) a. *[...] the Author first gives you the **Analysis or Algebra, whereby** all his General Methods of finding two Means were invented. (NA, An Account of Some Books, 1669)*
b. *[...] the Author first gives you the **Analysis or Algebra, by which** all his General Methods of finding two Means were invented.* (generated alternative)

According to Biber et al. (1999, p. 82), this analytic version is used in RCs with adverbial gaps and "[c]ommon only in academic prose, especially *in which* and *to which*". While they are perfectly common in German, PAs are relatively rare in English. Consequently, there are only a few studies reflecting on the specific use of pronominal adverbs in the relativizer position. PAs form a diverse group of compounds consisting of an adverbial element (*where, here, there*) and a prepositional one (*in, on, by*, etc.). Their formation process, adverbialization, is common to all Germanic languages and is one stage of grammaticalization (cf. Österman, 1997, p. 191). While PAs consisting of *there-* and *here-* + preposition are used as referential pronouns, PAs consisting of *where-* + pronoun can be used as either interrogative or relative pronouns. Despite being used only very rarely nowadays, PAs did represent a non-negligible subgroup of relativizers in the early days of the Late Modern Period's specialized discourse. While the first forms of PAs date back to Old English (Österman, 1997), PAs, as we know them today (i.e. *whereby, therefore, hereby*, etc.), originate in Middle English, a period of experimentation and composition (Mellinkoff, 2004; Österman, 1997). From this period on, their productivity, as well as frequency, first increases towards the Middle English Period and declines afterwards in types and overall frequency. Nevalainen

& Raumolin-Brunberg (2012, p. 203) describe this downward trend as typical of the general "typological drift [of English] from synthetic to analytic" proposed by (Sapir, 1921, p. 165–168), where a formerly synthetic form such as *whereby* is replaced by an analytic prepositional phrase, i.e. by *which/it/that*.

In RCs, PAs function as discourse connectives (cf. Nedoluzhko & Lapshinova-Koltunski, 2018) and through their composition of a poly-functional pronominal component (deictic, relative, indefinite, interrogative, and negative) and a prepositional component, convey a meaning of circumstance (place, time, manner) (cf. Ludiková, 1987). PAs are "multi-functional cohesive devices" (cf. Nedoluzhko & Lapshinova-Koltunski, 2018). Tracing *there-* compounds diachronically from EmodE to PDE, Österman (1997) finds that PAs take on increasingly abstract meanings on their way to grammaticalization, representing "the most grammaticalized end of the adverb category". For instance, the prepositional part in *therein* at earlier stages was used in a rather local sense, then gradually referred to more abstract circumstantial aspects. Being multi-functional, yet dense cohesive devices, pronominal adverbs have received special attention in jurislinguistics (linguistics of the language of the law) and are described to have "gradually disappear[ed] from general use" (Osminkin, 2020, p. 58) and instead settled in legal and religious texts (Crystal, 1969; Tiersma, 1999; Williams, 2007). Also, Mellinkoff (2004) mentions their diachronic integration in the language of the law and Österman (1997) points to their primary association with formal genres. In PDE, PAs give a "formal, old-fashioned, archaic, literary touch" (Sinclair, 1995). In line with that, promoters of the Plain English movement – Wydick & Sloan (2005); Williams (2007); Adler (2012) as well as Tiersma (1999, p. 96, p. 204) – consider pronominal adverbs in English "anachronisms that reduce understanding" (Osminkin, 2020, p. 59) and should not be used outside the language of the law. However, the perception of their linguistic efficiency diverges. Chovanec (2012, p. 2) regards PAs as "an efficient means of constructing cohesion and cross-referencing" by avoiding "undesirable repetitions in legal texts" (Osminkin, 2020, p. 59).

The existing research on relativizer choice during the lModE period has shown that the standard relativizers have not changed in terms of the inventory, but that their distributions in different genres have changed. However, there is no research in terms of relativizers and their syntagmatic contexts. Regarding the "non-standard" portion of relativizers consisting of pronominal adverbs, we have learned that they have gone through a process of decline and possibly "splitting" into their pronominal component and the relativizer *which*. The claim that the relativizer inventory has not changed in its scope should therefore be revisited so as to account for PAs as members of the paradigm.

2.3.4.2 German

Most studies on German relativizers (Ebert, 1986; Ebert et al., 1993; von Polenz, 1991; Ágel, 2000; Fleischer, 2013; Brooks, 2006; Dal & Eroms, 2014; Pickl, 2020) exclusively 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. Similar to English, in German there are also two major relativizers competing with each other, i.e. *d.** and *welch.**. The two types have a similar history to the relativizers *that* and *which*. *D.** is the older relativizer of the two originating in Old High German (OHG, 750–1050) (cf. Pickl, 2020, p. 245). While in OHG, *d.** was often used together with relative particles such as *de-* or *dar-*, today it is used alone. *D.** is the most frequent variant in PDG; however, at the end of the 16$^{\text{th}}$ c., *d.** met with the serious competition of *welch.** (Brooks, 2006, p. 135, cited by Pickl, 2020, p. 245).

Similar to *which*, *welch.** joined the relativizer system later than *d.**. Ebert et al. (1993, p. 446), for instance, assume that *welch.** found its way into the German relativizer system by way of Dutch in the 13$^{\text{th}}$ and 14$^{\text{th}}$ c. It was first integrated into Low German and from there spread to High German in the 15$^{\text{th}}$ c. (Brooks, 2006, p. 123, cited by Pickl, 2020, p. 245). Being established in the German language, towards the 16$^{\text{th}}$ c., *welch.** was increasingly used in rather formal and academic contexts as reported in historical grammars (Ebert, 1986, p. 446); (Ebert et al., 1993, p. 161). Possibly due to a sharp increase in (written) scientific literature production, at the end of the 16$^{\text{th}}$ c. the use of *welch.** also experienced a sharp increase (cf. Brooks, 2006, p. 135). Dal & Eroms (2014, p. 241) even claim that *welch.** in the 19$^{\text{th}}$ c. almost replaced *d.** in written German before declining in frequency again. However, results of a frequency analysis by Pickl (2020, p. 249 ff.) based on the corpora provided by the Digital Dictionary of the German Language (DWDS) show that the frequency of *welch.** in general has always been far below the frequency of *d.** with an especially sharp decrease in the second half of the 19$^{\text{th}}$ c., possibly due to a normative suppression (Ágel, 2000, p. 1883). In terms of its distribution in the different registers covered in the DWDS (Fiction, Science, Newspapers, LSP), Pickl (2020) finds some interesting differences: the decline first started in fictional texts, while scientific texts used *welch.** most frequently and for the longest time. Indeed, in scientific texts, the frequency of *welch.** even surpasses the frequency of *d.** at some points (between 1720 and 1740 and between 1850 and 1880). However, *welch.** declines dramatically after 1880. This finding is in line with von Polenz (1999, p. 5) stating that the general use of *welch.** declined remarkably in the 18$^{\text{th}}$ and 19$^{\text{th}}$ c. In line with its decline in numbers, according to Pickl (2020, p. 245, cites Duden 2011), *welch.** is nowadays described as "stylistically marked as formal and unwieldy" and "as an educated written variant unfamiliar in the dialects and in colloquial language" Pickl (2020, p. 245, cites Von Polenz, 1999, p. 356). Pickl (2020,

p. 246) also mentions the normative influence of 17th and 18th c. grammarians such as Georg Schottelius (Schottel, 1612–1676), Johann Christoph Gottsched (1700–1766), and Johann Christoph Adelung (1732–1806). Interestingly, the three grammarians evaluate the hierarchical order of the relativizers differently. For Schottel, the main relativizer is *d.**, while *welch.** functions as an alternative "where it would stand next to a homonymous form of the definite article in order to avoid repetition" (Schottel, 1967, p. 700 f., cited after Pickl, 2020, p. 247). Almost 100 years later, the evaluation changed. For Gottsched (1748, p. 237, cited after Pickl 2020) only *welch.** is a 'proper' relative pronoun and *d.** may also be counted as a relative pronoun, resulting in the inverse order of the hierarchy. Finally, Adelung (1783, p. 711) agrees with Gottsched, evaluating *welch.** as "the most complete relativizer". He also makes register-specific assessments of the relativizers and describes *welch.** as being more appropriately used "'in solemn speech' ('in der feyerlichen Rede'), whereas *d.** is used in 'the private/familiar way of writing' ('der vertraulichen Schreibart')". Pickl (2020, p. 247) thus concludes that "In the works of these three grammarians, we find a development of increasing explicitness as regards the acceptability of the individual forms and their stylistic assessment".

In German, PAs (also called prepositional adverbs in some German grammars) can consist of *da-*, *hier-* and *wo-* as the first constituent and a preposition as the second, and historically result from grammaticalization processes joining the formerly separate parts to a grammatical unit (Pittner, 2008). In German, the combinations of *wo(r)* + preposition (w-PAs) can be used as relativizers (as well as interrogative pronouns). PAs can functionally be described as pronouns since they can be used as anaphora, cataphora or have a deictic function. Syntactically, Pittner (2008) describes them as *Pro-PPs*, since they can practically take any function that a full prepositional phrase (PP) can take. Negele (2012) classifies PAs syntactically as relative, interrogative, phoric and deictic adverbs. PAs in German can only be formed with the "older, more grammaticalized prepositions, the so-called 'primary prepositions', such as *an*, *auf*, *aus*, *bei*, *durch*, *für*, *gegen*, *hinter*, *in*, *mit*, *nach*, *neben*, *ob*, *über*, *um*, *unter*, *von*, *vor*, *wider*, *zu*, *zwischen* and *trotz*, *wegen* and *während* (Helbig & Buscha, 2001, p. 353). The latter three, however, can only attach to *dem*, *des/dessen*, while all the others can theoretically be combined with *wor-* and be used as relativizers. According to Pittner (2008), PAs existed in both Old English and Old High German and the formation of PAs was productive in German for much longer than in English. In German, the last new creation of PAs was made in the 17th c., while in English, the productivity of PAs ended in the 14th c., which can be derived from the fact that no preposition created after 1300 is part of a PA (cf. Müller, 2000, p. 173). The most striking difference between PAs in English and German seems to lie less in their productivity than in their use. While in German, w-PAs occur in interrogative and relative clauses, in English wh-PAs are extremely rare. Pittner (2008) attributes this to the fact that English prefers preposition stranding in cases where in German a PA would be used.

(14) a. *the material* ***that*** *it consists* ***of***
b. *das Material,* ***woraus*** *es besteht*

While English has a general preference for the *analytic* version (relativizer + preposition), when the preposition is stranded, or the preposition is directly followed by the relativizer in a pied-piping construction, in German, the interchangeability of the analytic and the synthetic forms (PAs) is semantically motivated by the animacy of the head noun (Fleischer, 2002, p. 23), i.e. PAs can only be used as referring expressions for inanimate PPs (Example (15)):

(15) a. *Das Ding, womit das Kind spielte.*
b. **Die Frau, womit das Kind spielte.*

In German, PAs are furthermore especially frequent as means of extended reference to a predicate (Example (16-a)). In such cases, the use of a PA is obligatory and a replacement by an analytic combination of preposition + relative pronoun is not licensed (Example (16-b)).

(16) a. *Sie* ***gewann*** *den ersten Preis,* ***worüber*** *sie sich sehr freute.*
b. *Sie* ***gewann*** *den ersten Preis,* ***über das*** *sie sich sehr freute.*

While PAs in present-day English are stylistically regarded as archaic, formal and used (if at all) in legal language, in German, PAs are perfectly common and do not have a clear-cut stylistic affiliation.

2.4 Summary of Part I

In Section 2.1 of the present chapter, we started out by defining the core concepts building the base for our analyses: *complexity*, *efficiency*, and *utility*. We then examined the main extra-linguistic factors shaping the development of scientific English and German (Section 2.2): the institutionalization of science, standardization of the vernaculars, and scientific publishing practice in the vernaculars. In Section 2.3, we gave an overview of general linguistic developments as well as developments specific to the scientific meta-register between 1650 and 1900 in English and German. We finally presented extant work on the diachronic development of RC usage in scientific writing.

We would now like to summarize our insights into complexity and what we have learned about the extra-linguistic and linguistic factors in the development of scientific writing, and from this, generate hypotheses for the present thesis. The resulting hypotheses will guide our corpus studies looking at the development of lexicogrammatical (Part III) and syntactic (Part IV) complexity in scientific writing over time.

In Section 2.2, we have seen that developments regarding extra-linguistic factors across English and German were similar in terms of the institutionalization of scientific work and publishing practice. Both the UK as well as the German-speaking area had scientific academies, the German *Academia Naturae Curiosum* and the British *Royal Society*. Both academies also published their own scientific journals: the Royal Society's *Philosophical Transactions* and the Leopoldina's *Miscellanea Curiosa*. Also, for both languages, the efforts toward standardizing the vernaculars were at their peak in the 18$^{\text{th}}$ c. Thus, English and German had equal opportunities to develop as scientific languages in terms of institutionalization of scientific work and publishing. In terms of language policy, German was even further ahead than English, as it had an official language academy called "*die Fruchtbringende Gesellschaft*" founded as early as 1617, while English had no such institution at that time.

The key distinction between scientific writing in German and English lies in the shift from Latin to the use of vernacular languages. The *Royal Society* advocated for scientific communication in the vernacular from its inception, but German scientific institutions and publication practices continued to use Latin well into the 19$^{\text{th}}$ c. Furthermore, it can be inferred that the formation of a nation-state played a role in the adoption of vernacular languages in scientific writing. Also in this respect, the UK was ahead of Germany. The UK became a unified nation-state in the early 18$^{\text{th}}$ c., while Germany did not become a nation-state until the late 19$^{\text{th}}$ c. The UK's social and cultural advantage over Germany was also evident in the Industrial Revolution, which is widely regarded as a pivotal moment in human history and a significant catalyst for innovation and progress.

The socio-cultural changes discussed above are also reflected in the linguistic development of the scientific meta-register (discussed in Section 2.3). The existing literature on the history of scientific writing in English and German indirectly highlights the most significant difference between the two languages. While there is extensive research analyzing the development of scientific writing in English, based mainly on the *Royal Society's Philosophical Transactions*, there is practically no equivalent quantitative work covering the 17$^{\text{th}}$ and 18$^{\text{th}}$ centuries in German. This is because German was not yet used for scientific communication in an institutionalized manner. We will now combine what we do know about the actual linguistic features of scientific writing in both languages, with our insights into *lexico-grammatical and syntactic complexity* as we defined them in Section 2.1.3.3, Figure 2.6.

Lexico-grammatical complexity can become manifest in terms of the variability of *syntagmatic* and *paradigmatic* contexts of words. Previous research on the *lexico-semantic* level of English scientific writing has shown a trend of lexical innovation in response to the need for new terminology, as well as a move towards specialization. This is reflected in nominal patterns that exhibit increasing information density and become less predictable in their syntagmatic contexts over time. Similarly in German, there seems to have been an expansion of the vocabulary, with many Latin words being integrated into the German scientific language. At the *lexico-*

grammatical level (i.e. function words), however, there seems to have been a trend toward conventionalization in usage leading to a decrease in *paradigmatic richness*, i.e. a limited range of linguistic choices in a particular context. These findings are based on information-theoretic measures such as surprisal and entropy taking syntagmatic and paradigmatic context into account. Since entropy and surprisal are complexity measures that correlate with the cognitive effort required for expectation-based processing, it appears that processing-related complexity plays a significant role in the formation of scientific writing registers. Previous work related to lexico-grammatical complexity suggests a trade-off between the productivity of lexico-semantic choices and the convergence of lexico-grammatical options to a reduced set, and we assume that this is in order to minimize processing complexity overall and hence to serve the utility of scientific writing for expert-to-expert communication. Previous work on RC usage over time has also shown that a similar trend of grammatical consolidation seems to be at work for relativizers. Both in English and in German, the set of relativizers seems to become conventionalized with the assignment of specific relativizers to certain contexts of usage, adapted to fulfilling their function in scientific communication. However, the latter insights are merely based on frequency distributions of relativizers in different registers over time. Our definition of *lexico-grammatical* complexity (2.1.3, Figure 2.6) instead is based on the assumption that complexity is inherently processing-related and thus has to be approximated by means of measures that are known to be correlated with processing effort. Therefore, in order to study the evolution of the lexico-grammatical complexity of RCs in scientific writing in English and in German, we employ information-theoretic metrics that have been linked to expectation-based processing effort. Entropy may thus serve as a valuable measure to give a processing-based account of lexico-grammatical reduction in *paradigmatic richness* of the relativizer paradigm leading to lower expectation-based processing cost over time. Taking into account the syntagmatic context of a linguistic unit, surprisal may serve to detect increasingly conventionalized and thus predictable usage contexts of RCs, contributing to lower processing effort on the syntagmatic axis of linguistic structure.

The existing work regarding ***syntactic complexity*** in the scientific meta-register shows a trend toward syntactic compression in terms of a preference for nominal over clausal renderings. For both languages, we have seen that in scientific discourse, subclauses, including RCs, become less frequent over time. However, RCs are still the most frequent subordinate clause type in scientific writing. Overall, this syntactic compression is held to be favorable for efficient specialized expert communication. Previous work on the diachronic development of *syntactic complexity* in scientific writing is, however, either descriptive (as is the case for German) or frequency-based. While there have been made assumptions as to how the shift from clausal to phrasal complexity might affect processing (i.e. lower explicitness resulting in higher difficulty for non-expert readers), the frequency-based findings have not been explicitly associated with measures known to be correlated with memory-based processing effort

such as dependency length and expectation-based indicators such as accessibility. In Section 2.1.3.3, we defined *syntactic complexity* as being modulated by the composition of syntactic dependencies in length (DL) in general and in type of RCs (i.e. accessibility) in particular. Applying our definition of *syntactic complexity* to what we have learned about scientific writing, we believe that syntactic compression on the clausal level is tightly connected to the length of syntactic dependencies. To trace the development of *syntactic complexity*, DL may thus serve as a valuable measure to account for a trend towards syntactic compression in scientific writing while effectively accounting for memory-based processing effort. While we have not found any evidence for the modulation of accessibility influencing the complexity of RCs in scientific writing over time, accounting for the distributional developments of RC extraction types diachronically can give valuable insights on whether and to what extent expectation-based processing effort is involved in syntactic complexity of RCs in Late Modern English and New High German.

Chapter 3

Hypotheses

The general assumption underlying this thesis is that scientific language becomes grammatically less complex and therefore more efficient for scientific communication. We will trace this development by looking at RCs, which represent grammatically complex and explicit ways of nominal post-modification. In this chapter, we will state our hypotheses about the development of RCs in scientific language between 1650 and 1900 and delineate which measures of complexity (described in further detail in Chapter 5) we will apply to corroborate the generated assumptions.

3.1 Register formation

(1) **H1: RCs have contributed to decreasing grammatical complexity on different linguistic levels in scientific writing.**

The overarching hypothesis in this thesis is that scientific writing over time develops toward lower grammatical complexity (as defined in Section 2.1.3) in order to counterbalance pressures deriving from an expansion on the lexico-semantic level (new incoming vocabulary) and increased lexical density. We assume that this trend toward lower grammatical complexity is specific to scientific writing as compared to general language, and we will test this claim by comparing corpus findings from our scientific corpora (*Royal Society Corpus*, RSC and *Deutsches Textarchiv Wissenschaft Corpus*, DTAW) to findings from our general language corpora (*Corpus of Late Modern English Texts*, CLMET and *Deutsches Textarchiv General Corpus*, DTAG). We will trace register-specific complexity reduction on the lexico-grammatical and on the syntactic level by looking at the indicators of complexity discussed in Section 2.1.3. In the following sub-hypotheses, we will specify our assumptions about the two levels further.

3.1.1 Lexical complexity

While we know that on the lexico-semantic level, there is an expansion in scientific writing, we assume we will find a trend toward lower complexity and thus greater efficiency on the lexico-grammatical level in terms of *paradigmatic richness* and *syntagmatic predictability*.

3.1.1.1 Paradigmatic richness

We assume that the convergence of specific register-specific options to introduce RCs (i.e. relativizers) is a means of making communication more efficient through the avoidance of uncertainty about which relativizer will be chosen. At the same time, settling on a preferred option leads to the superfluity of other options which are expected to fade out of the picture. We specifically have pronominal adverbs in mind, which represent a synthetic form of *which* + preposition (e.g. *whereby*). The replacement of a large set of relativizers by analytic variants (e.g. *by which*) would lead to lower uncertainty about the specific relativizer chosen at the onset of an RC. We thus assume that

(2) **H1.1: scientific writing develops toward a reduction in paradigmatic richness as expressed by lower entropy indicating the conventionalization *of* and the lower uncertainty *about* the relativizer choice.**

We test this assumption in Chapter 6.

3.1.1.2 Syntagmatic predictability

Scientific writing becomes increasingly conventionalized mainly on the grammatical level. We therefore expect to find a trend toward conventionalized contexts in which RCs increasingly tend to occur. We thus assume that in scientific writing,

(3) a. **H1.2a: surprisal at the onset of RCs decreases over time,**
b. **H1.2b: surprisal of certain, preferred relativizers decreases over time indicating their higher predictability due to conventionalized contexts. Surprisal of less preferred relativizers increases over time due to less conventionalized contexts.**

We test these assumptions in Chapter 7.

3.1.2 Syntactic complexity

As discussed in Section 2.1.3.2, we look at three indicators of syntactic complexity: *intricacy*, *locality*, and *accessibility*. In the following, we will outline our hypotheses regarding each of these indicators.

3.1.2.1 Syntactic intricacy

Syntactic intricacy refers to syntactic embeddedness by means of subordinate clauses. Scientific English has been reported to become syntactically less intricate, i.e. showing fewer clausal embeddings over time. German scientific writing at first had a tendency to become more intricate until the early 19th c. by making increased use of hypotactic structures due to the piecemeal discovery of German syntactic possibilities, and only later this trend reversed toward an abandonment of intricate structures.

Apart from quantitative work on the usage of RCs over time, a diachronic reduction of the grammatical construction of RCs can be supported by the efficiency principle *Minimize Forms* suggested by Hawkins (2004), which predicts that if something can be expressed with a short and a long form, it will be expressed in the shorter form. For an RC, this means that only RCs which do not have a shorter alternative will be built, and those that do have shorter alternatives will disappear. For scientific writing in particular, this is especially plausible, since RCs serve as highly explicit ways to further define a head noun. As new concepts (e.g. discoveries, chemical compounds, etc.) are created, they first have to be explicitly defined before being established later on in the general knowledge of the community. As soon as a concept is sufficiently defined and established, the form can be reduced to a minimum and an RC is not necessary anymore. Relying on these previous observations, we assume that

(4) **H1.3a: RCs overall will become less frequent in scientific writing.**

Apart from the general abandonment of superfluous RCs, we also assume that RCs will occur in a less accumulated way, i.e. fewer RCs will occur in one sentence on average. Hence,

(5) **H1.3b: the number of RCs within a sentence will decrease on average in scientific writing.**

While the first hypothesis looks at the overall number of RCs found in scientific language, the second one concentrates on the degree of embeddedness in those sentences containing at least one RC. The more RCs a sentence contains, the more complex and harder to process the sentence becomes. We test our hypotheses in Chapter 9. Since RCs represent clausal subordination creating long-distance dependencies, the complexity created by the frequency and density of RCs (per sentence) can be attributed to the cognitive mechanism of working memory. This leads us to the next hypothesis, which regards *locality* as another indicator of syntactic complexity.

3.1.2.2 Locality

RCs represent clausal embeddings creating relatively long dependency relations between the head noun and the embedded verb of the RC (see Figure 3.1).

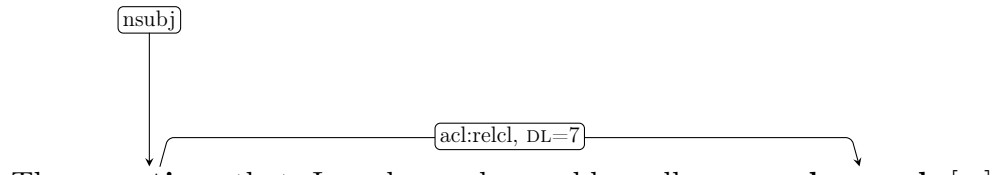


Figure 3.1: Dependency length created by a longer RC.

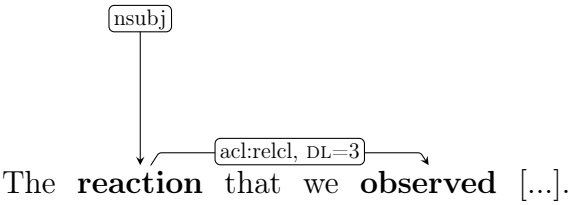


Figure 3.2: Dependency length created by a shorter RC.

Since RCs overall are expected to become less frequent in scientific language, it can be assumed that this reduction of RC frequency and density leads to an overall minimization of average dependency length (ADL, introduced in Section 2.1.3.2 and described in Section 5.2.2) in scientific language. Following the *Dependency Locality Theory* (DLT, Gibson, 2000) stating that processing effort depends on the distance between two syntactically related elements (head and dependent), we assume that, also within RCs, DL should become shorter to optimize memory-based processing effort (compare Figure 3.2).

(6) **H1.4a: Scientific writing develops toward shorter ADL leading to greater locality overall.**

(7) **H1.4b: Scientific writing develops toward shorter DL within the construction of RCs, i.e. between the head noun and the embedded verb of the RC.**

We test our hypotheses in Chapter 10.

3.1.2.3 Accessibility

The Accessibility Hierarchy (Keenan & Comrie, 1977; Keenan & Hawkins, 1987) predicts that across languages, an RC type is more difficult to process the lower down the hierarchy it is extracted from (compare Equation 2.1 in Section 2.1.3.2). Thus, subject RCs are assumed to be easier to process than object RCs, etc. Hence, we assume that

(8) **H1.5: in scientific writing, over time, more accessible RC types (i.e. subject RCs) will be preferred over less accessible RC types.**

We test this hypothesis in Chapter 11.

3.2 Language-specific contrasts

Our second overarching hypothesis is that we will not only find differences when comparing the scientific meta-register to general language, but that we will also find differences comparing the two languages German and English. Based on the extra-linguistic and linguistic historical differences in the development of the two meta-registers, we specifically expect to find a *time-shifted* development in German scientific writing as indicated by

(9) **H2: a later turn toward lower grammatical complexity in scientific German compared to scientific English and thus a more linear development toward lower grammatical complexity in English, while scientific German will first increase in complexity until the early 19th c. and decrease afterward.**

Part II

Data and Methods

Chapter 4

Corpora

In the present Chapter, we describe the corpora used in our analyses. Since we are interested in the development of the scientific meta-register in English and German, we have two scientific corpora covering almost the exact same time span, i.e. 1650–1899: the Royal Society Corpus (Fischer et al., 2020, RSC) for English and the DTAW, which is the scientific portion of "Deutsches Textarchiv" (Geyken et al., 2018, DTA) for German. Note that the English corpus is compiled of texts from the "Philosophical Transactions" and "Proceedings of the Royal Society of London" starting in 1665. The German texts are derived from the "Deutsches Textarchiv", which provides texts even from earlier time periods. However, to establish comparability between the corpora, we only consider texts between 1650 and 1899 in the DTA. Furthermore, to be able to evaluate whether our observations are specific to scientific writing, we also use two "general" language corpora for comparison: the CLMET (Diller et al., 2011) for English, and for German the DTAG, which is composed of the general language portion of the DTA.

4.1 The English corpora

4.1.1 The RSC

For scientific English, we use the Royal Society Corpus (RSC_6.0_Open, Fischer et al., 2020). The corpus covers over 250 years of scientific texts taken from the *Philosophical Transactions* and *Proceedings of the Royal Society of London* between 1665 and 1920. The complete version contains 17,520 texts and 78.6 million tokens. The corpus is annotated with a standard linguistic annotation such as parts of speech (using the Penn Tag Set, Santorini, 1990), as well as surprisal[1]. The historical word

[1]For a detailed description of the calculation and the underlying rationale, please refer to Section 5.1.2.

forms were normalized using VARD (Baron & Rayson, 2008). Tokenization, lemmatization, and POS tagging were done with TreeTagger (Schmid, 1994). The corpus also contains structural annotation regarding the topic(s) covered by each text. Figure 4.1 shows the distribution of topics covered in the RSC. The topics were determined with topic modeling (Blei et al., 2003) using MALLET (McCallum, 2002) as described by Bizzoni et al. (2020). Note that for this thesis, only the years 1665–1899 were taken into account. Table 4.1 shows the number of texts, tokens, and sentences in this time span grouped by 50-year periods. For our dependency length analyses, we created another, slightly reduced version of the corpus, whose creation including the parsing process and necessary preprocessing steps we describe in detail in Section 4.3.

Years	# Texts	#Types	# Tokens	#Sentences
1665–1699	1 325	89 823	2 582 856	74 709
1700–1749	1 686	98 175	3 414 795	120 238
1750–1799	1 819	170 691	6 342 489	208 125
1800–1849	2 774	269 551	9 112 274	333 632
1850–1899	6 754	843 220	36 993 412	1 770 027
Total	14 358	1 204 294	58 445 826	2 506 731

Table 4.1: RSC corpus statistics.

4.1.2 The CLMET

For "general" English, we use the Corpus of Late Modern English texts (CLMET, De Smet, 2006). The corpus covers over 200 years (1710–1920) of text from different registers (De Smet, 2006; Diller et al., 2011) and represents a collection of public domain texts obtained from online archives (Oxford Text Archive and Project Gutenberg). The register mix contains narrative fiction, non-fiction, drama, letters, and treatises. The complete size of the corpus is smaller than that of the RSC, with around 40 million tokens and approx. 350 texts. The corpus was annotated with the same tools as the RSC. For this thesis, again, only the years 1700–1899 were taken into account. Table 4.6 shows the number of texts, tokens, and sentences in this time span grouped by 50-year periods.

Years	# Texts	#Types	# Tokens	#Sentences
1700–1749	26	51 570	3 668 435	126 908
1750–1799	90	118 718	10 826 761	387 257
1800–1849	71	136 680	10 987 325	410 373
1850–1899	90	135 057	10 772 463	466 682
Total	277	442 025	36 254 984	1 356 016

Table 4.2: CLMET corpus statistics.

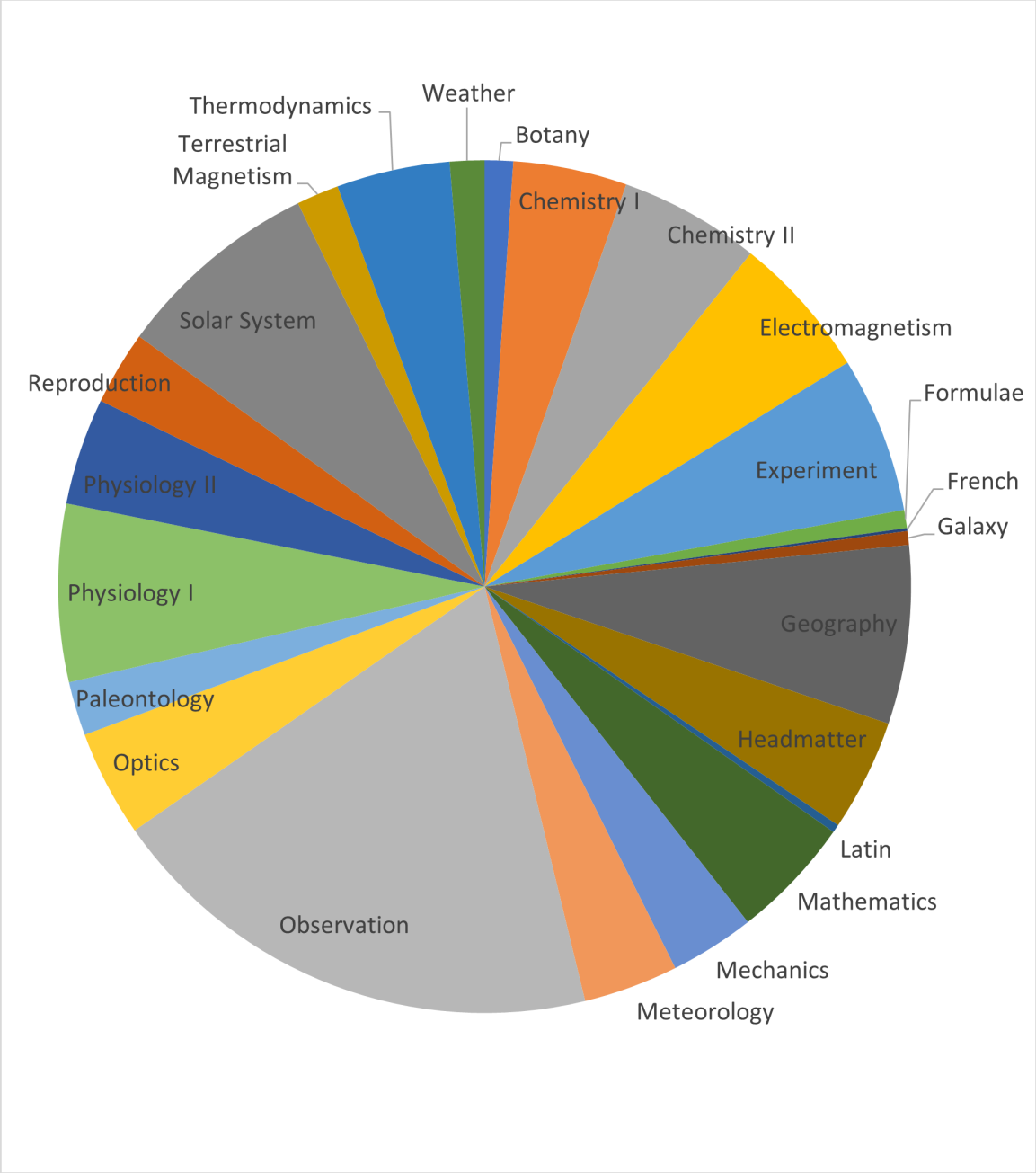


Figure 4.1: Scientific topics covered in the RSC.

4.2 The German corpora

For German, all texts from 1650–1900 are retrieved from the *Deutsches Textarchiv* (Geyken et al., 2018, DTA). The DTA is a linguistically annotated full-text corpus freely available online. It was compiled at the *Berlin Brandenburgische Akademie der Wissenschaften* (BBAW). The DTA contains texts from four different genres (narrative fiction, newspapers, non-fiction, scientific texts) from the time between 1650 and 1900. The choice for each text to be included was made according to linguistic and lexicographic aspects (Geyken et al., 2018) under the premise of balancing the distribution of texts with regard to the different scientific disciplines and genres[2]. The DTA contains seminal texts within each genre, which are held to be influential and well-received works representative of German literature, and in the case of the scientific texts representative of the development of scientific disciplines (Geyken et al., 2018). On the basis of the full version of the DTA, we build our ***scientific corpus*** (**DTAW**) including all scientific texts from the full DTA and our ***general corpus*** (**DTAG**) including the texts from all other genres in the DTA. Both portions contain metadata (e.g. author, publication year, title, etc.) and linguistic annotation (e.g. tokens, lemmas, normalization, parts of speech), which were retained from the original format provided by the BBAW. The German part-of-speech (POS) annotation is based on the "Stuttgart-Tübingen Tagset" (STTS, Hinrichs et al., 1995). The DTA comes with canonicalized wordforms created with CAB (Jurish, 2011)[3]. The tokenization of the DTA was created using the specifically built tool DTA-Tokwrap (Jurish, 2011)[4]. Further annotations, which we added, include surprisal (Section 5.1.2) and dependency length (Section 5.2.2).

4.2.1 The DTAW

The corpus size of this portion of the corpus is approx. 80 million tokens. Figure 4.2 shows the distribution of scientific disciplines covered in the DTAW. The disciplines were manually assigned according to recommendations by members of the BBAW specialists in the respective disciplines. As sources, only first editions were used to ensure representativeness of the respective historical linguistic stage. Comparing the composition of topics in the RSC represented in Figure 4.1 to the composition of disciplines in the DTAW (Figure 4.2), it is obvious that the two corpora are quite different. While the RSC exclusively contains *natural* scientific topics, the sub-disciplines in the DTAW cover all kinds of scientific areas (i.e. humanities, law, natural science, etc.). We need to bear this in mind in the interpretation of results.

[2]Translated from (Berlin Brandenburgische Akademie, 2020).
[3]Demo available at (Jurish, 2012).
[4]Documentation available at (Jurish, 2020).

Years	# Texts	#Types	# Tokens	#Sentences
1650–1699	160	282 329	6 680 789	219 621
1700–1749	189	361 211	9 176 183	370 846
1750–1799	337	493 694	16 902 262	702 128
1800–1849	331	453 485	14 632 047	530 533
1850–1899	352	906 123	31 772 335	1 185 717
Total	1 369	1 814 309	79 163 616	3 008 845

Table 4.3: DTAW corpus statistics.

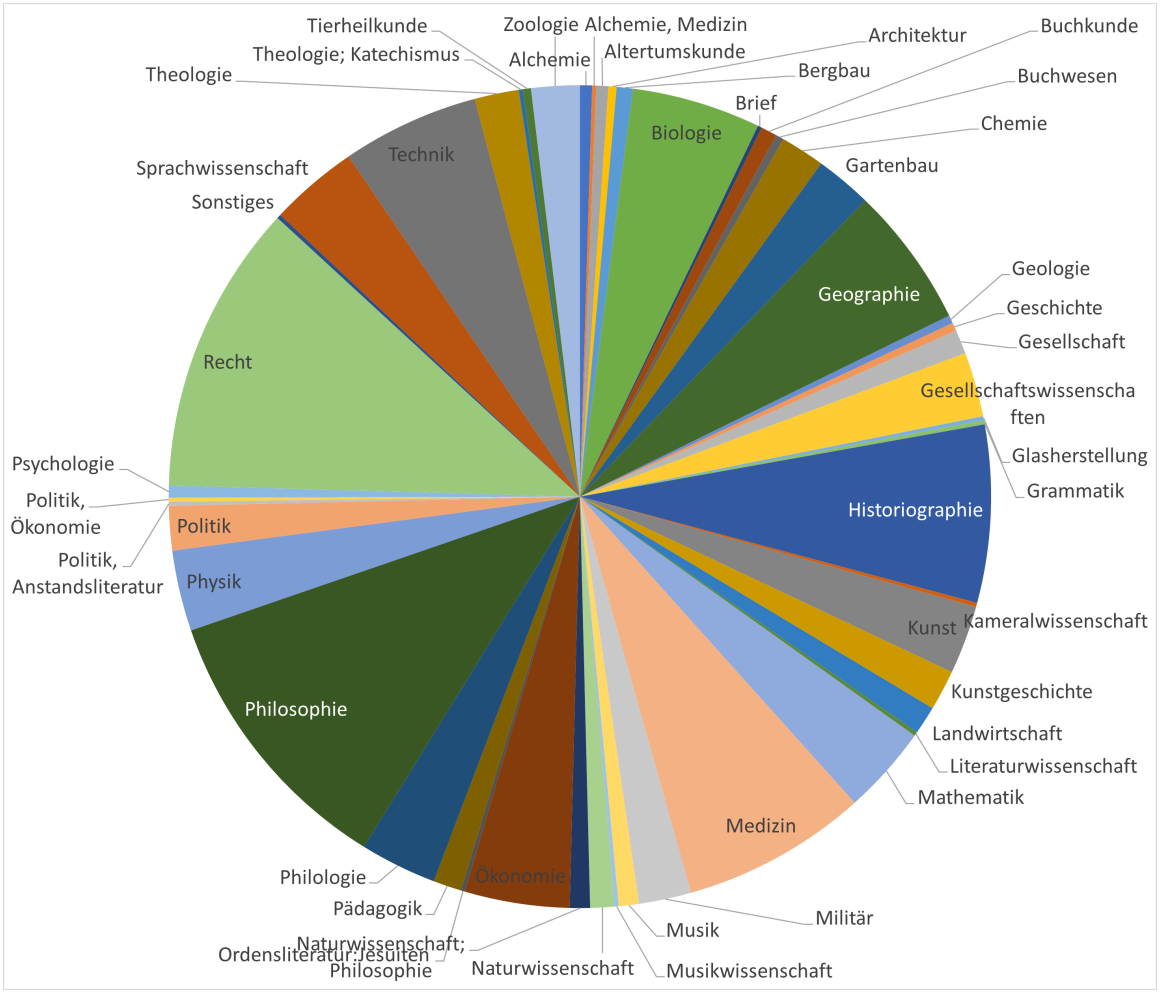


Figure 4.2: Scientific disciplines covered in the DTAW (subcategories).

4.2.2 The DTAG

"General" German is represented with approximately 60 million tokens including non-fictional (topics: politics, society, pedagogy, mathematics, zoology, miscellaneous, regulations, technology, popular scientific texts) and fictional prose texts (poetry, drama, prose, autobiography, travel literature, novels).

Years	# Texts	#Types	# Tokens	#Sentences
1650–1699	110	500 936	13 806 389	586 924
1700–1749	122	380 614	14 130 208	533 695
1750–1799	179	318 269	11 315 902	528 137
1800–1849	200	365 410	12 331 892	527 444
1850–1899	142	373 120	10 743 945	526 074
Total	753	1 316 607	62 328 336	2 702 274

Table 4.4: DTAG corpus statistics.

4.3 Syntactic annotation

To analyze scientific writing in terms of its syntactic complexity, we need to add syntactic information such as syntactic dependencies and dependency length (DL) to our corpora. In the upcoming section (Section 4.3.1), we will discuss various preprocessing steps[5] necessary to prepare the data for syntactic annotation (parsing). These steps are crucial to achieve the best possible parsing quality and minimize errors associated with historical language data. The non-digital origin of historical data requires significant preprocessing efforts, which involve tasks such as standardizing data formats, correcting OCR errors, and annotating metadata. These measures are necessary to ensure the accuracy and reliability of the data, which may contain variations in spelling, morphology, and syntax (Menzel et al., 2021), thereby presenting challenges to linguistic annotation and analysis. Typical linguistic annotation bottlenecks such as variations in spelling, morphology, and syntax, especially in word order, can hamper the parsing process. In particular, incorrect sentence splitting is a significant issue that negatively impacts syntactic parsing (Juzek et al., 2019a,b). In Section 4.3.2, we will describe the parsing process and, in Section 4.3.4, report on the manual evaluation of the syntactic annotation by linguistic specialists. We will also compare the parsing accuracy of preprocessed data to non-preprocessed data to determine if preprocessing significantly improves parsing quality.

4.3.1 Preprocessing

In the subsequent sections, we will present the parsing pipeline, which we specifically designed to accommodate the unique characteristics of our historical corpora while ensuring their cross-linguistic comparability. Notably, our corpora already contain essential linguistic annotations based on customized processing. Therefore, we aim to retain as many of the existing linguistic annotations as possible to facilitate the parsing process. As part of our preprocessing efforts, we begin by normalizing the German corpora for punctuation. Specifically, we replace the previously prevalent

[5]Most of this process as well as a description of the resulting scientific corpora DTAW_UD-parsed_1.0 and RSC_UD-parsed_1.0 is described in (Krielke et al., 2022).

virgule (slash, see Example (1-a)) with the corresponding comma (as shown in Example (1-b)).

(1) a. *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)
b. *Wann jemand etwas seinem Nächsten zum Besten aufrichtig heraus gibt, so gering es auch ist, billig zu Dank soll angenommen werden.*

Furthermore, before parsing our corpora with the UDpipe Parser (Straka, 2018), we applied several rules to extract "good sentences" (**GS**) only, i.e. sentences that fulfil specific requirements that a parser normally "expects". By feeding only sentences which were controlled for their well-formedness to the parser, we expected to significantly improve the parsing results. To ensure this, we built on the preexisting annotation to detect non-sentential constructions as well as foreign-language sentences (*foreign*). Specifically, we deleted any sentence beginning with a word in lower case and the sentence preceding it (*incomplete*), sentences with less than eight tokens (*too short*), as well as sentences lacking a verb (*verbless*). To exclude foreign-language sentences, we ran the language recognizer LangID (Lui & Baldwin, 2012) on each sentence in the corpora and excluded all sentences in languages other than English or German respectively. The preprocessing was only applied to the RSC, DTAG, and DTAW since those were found to contain several sentences representing difficulties (especially end-of-sentence errors) for a parser. The CLMET did not exhibit this problem, which is why it was excluded from our preprocessing. After preprocessing, we obtained approximately 26 million tokens for the scientific English corpus (RSC) and 74 million tokens for the scientific German corpus (DTAW) and 58 million tokens for the general German corpus (DTAG). For information on remaining tokens and sentences after applying the above rules, see Tables 4.5, 4.7 and 4.8. For a comparative evaluation (Section 4.3.4) of the improvement gained by the **GS** selection, we also retained all discarded "bad sentences" (**BS**).

4.3.2 UD-parsing

For the analysis of dependency lengths, we employed the framework of Universal Dependencies (UD), which expresses syntactic relations through dependencies: each element depends on another element, functioning as its head. In UD, in contrast to most other dependency frameworks, the head is taken to be the semantically salient element and the dependent modifies the head. The top-level head is the root of a sequence, which is typically the main verb of the matrix clause. For instance, in the sentence *There was a cat that ran away*, the relative pronoun *that* and *away* modify the embedded verb *ran*, which depends on the head noun *cat*, and the head noun depends on the root *was*. Figure 4.3 illustrates a dependency analysis of this example. To syntactically annotate the corpora, we used the UDpipe parser (Straka, 2018),

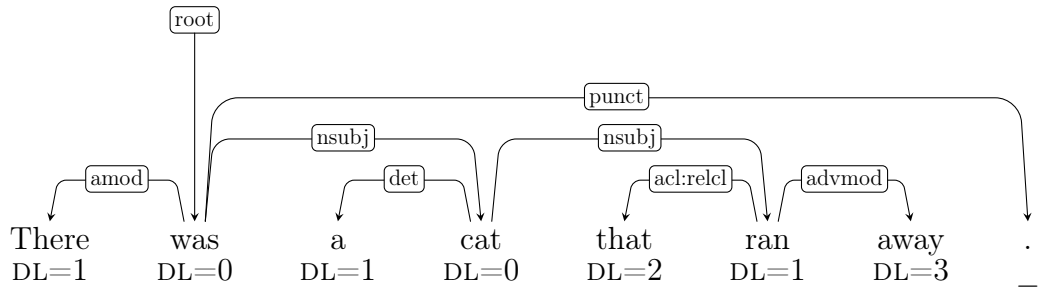


Figure 4.3: Graphic visualization of a simple sentence in the Universal Dependencies framework. The edges represent a dependency relation pointing from head to dependent; the numbers denote the dependency length (DL) between tokens.

which is based on the Universal Dependencies (UD) framework. The UD framework aims to be universal, i.e. suitable for all of the world's languages, and there are a great number of resources and tools available for postprocessing parsed output[6]. Importantly, UD-parsing labels nodes with syntactic functions such as *nominal subject, adverbial modifier*, etc. This is crucial for exploring the functions that are associated with dependency length minimization over time (Chapter 10).

Before parsing, the texts were extracted from the preprocessed corpora (now consisting only of GS) in such a way that metadata are preserved. We preserved the original sentence splitting and tokenization before passing the text to the parser. As the name suggests, the UDpipe 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 they should resemble our older historical data which still show more general language features presumably covered by a multi-genre model.
Once parsed, the German texts were augmented with one final, yet important preprocessing step. The German UD tagset does not include the `acl:relcl` tag, but only `acl`, which serves as an umbrella tag for any type of adnominal clause. Thus, to identify relative clauses, we further enriched the German resulting treebank with this information by applying the following rule: any token tagged as `acl` with a dependent whose POS tag is a relativizer (PRELS or PRELAT) should be renamed as `acl:relcl`.

Finally, we annotated the resulting treebanks with our measure of syntactic complexity, and dependency length (DL, described in Section 5.2.2). We calculated the DL of each token (excluding punctuation), the sentence length (SL) for each sentence, the summed DL for each sentence (SDL), and the average DL (ADL) per sentence and annotated them as positional attributes into the corpus, such that each token has the following information: token_id, word, lemma, pos, ud_label, ud_head, SL, DL,

[6]Documented at Universal Dependencies.org (2023b).

SDL, ADL[7]. The script for the extraction of "good sentences" is available on github[8].

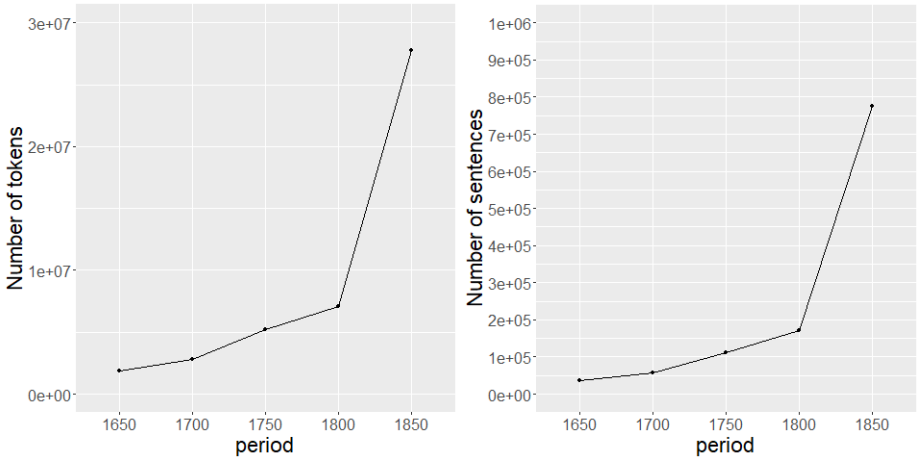
4.3.3 Corpus statistics

The following tables show the resulting numbers of tokens and sentences after preprocessing and parsing; the subsequent figures visualize the distributions of tokens and sentences across the 50-year periods.

4.3.3.1 RSC

Years	# Texts	# Tokens	#Sentences
1665–1699	1 207	2 196 793	42 238
1700–1749	1 658	2 860 204	57 747
1750–1799	1 816	5 205 741	112 844
1800–1849	2 709	7 260 221	177 181
1850–1899	6 586	28 310 228	810 990
Total	13 976	45 833 187	1 201 000

Table 4.5: RSC_UD-Parsed_1.0 corpus statistics.



(a) tokens **(b)** sentences

Figure 4.4: Number of (a) tokens and (b) sentences in scientific English (RSC) by 50-year periods.

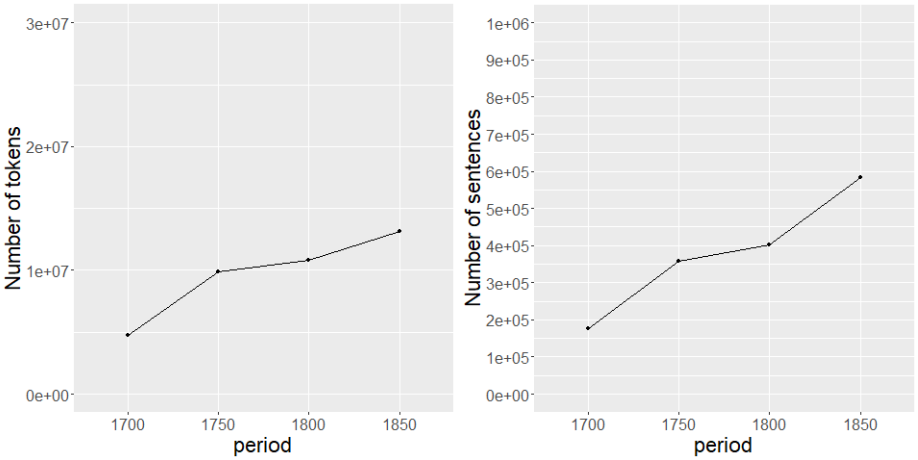
[7]The corpora resulting from preprocessing and enriched with UD parses are then called RSC_UD-Parsed_1.0, CLMET_UD-Parsed_1.0, DTAW_UD-Parsed_1.0 and DTAG_UD-Parsed_1.0. For the sake of brevity, when using the parsed corpora for our analyses of syntactic complexity, we will refer to them with their short names: **RSC**, **CLMET**, **DTAW** and **DTAG**.

[8]Created by Knappen (2022).

4.3.3.2 CLMET

Years	# Texts	# Tokens	#Sentences
1700–1749	26	4 931 584	177 079
1750–1799	90	10 265 464	357 176
1800–1849	71	11 221 759	402 025
1850–1899	90	13 743 371	584 020
Total	277	36 254 984	1 356 016

Table 4.6: CLMET (parsed) corpus statistics.



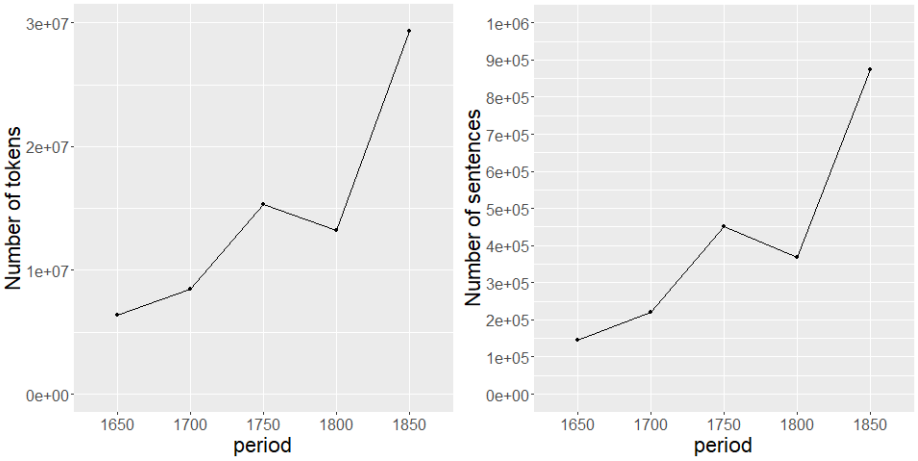
(a) tokens **(b)** sentences

Figure 4.5: Number of (a) tokens and (b) sentences in general English (CLMET) by 50-year periods.

4.3.3.3 DTAW

Years	# Texts	# Tokens	#Sentences
1650–1699	50	6 210 992	144 896
1700–1749	67	8 202 555	219 409
1750–1799	158	15 150 320	451 881
1800–1849	131	13 073 176	368 739
1850–1899	211	29 015 797	874 133
Total	617	71 652 840	2 059 058

Table 4.7: DTAW_UD-Parsed_1.0 corpus statistics.



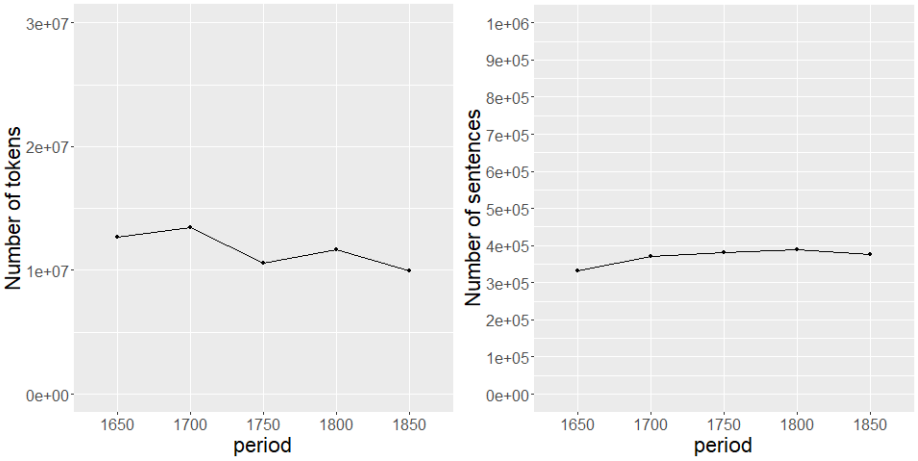
(a) tokens **(b)** sentences

Figure 4.6: Number of (a) tokens and (b) sentences in scientific German (DTAW) by 50-year periods.

4.3.3.4 DTAG

Years	# Texts	# Tokens	#Sentences
1650–1699	110	12 412 665	331 911
1700–1749	122	13 121 862	372 117
1750–1799	179	10 557 988	379 940
1800–1849	200	11 612 819	389 969
1850–1899	142	9 901 075	375 474
Total	753	57 606 409	1 849 411

Table 4.8: DTAG_UD-Parsed_1.0 corpus statistics.



(a) tokens **(b)** sentences

Figure 4.7: Number of (a) tokens and (b) sentences in general German (DTAG) by 50-year periods.

4.3.4 Parser evaluation

To evaluate the quality of the parses after the preprocessing steps described above, we sampled 100 sentences (20 from each 50-year period, e.g. 1650–1699) from the "good sentences" (**GS**) of the scientific corpora (RSC and DTAW) and evaluated them against 100 parsed sentences from those discarded by our filter (**BS**). The reason for evaluating only the scientific corpora was that due to their extensive use of tables and special formatting, the parses can be expected to come out worse than those of general language lacking these stumbling blocks. Furthermore, since we did not apply the preprocessing to the general English corpus, we would not have been able to compare the results against a non-preprocessed version. We trust that the insights obtained from the evaluation of the scientific texts can be generalized to the parsing quality of all four corpora. The samples were evaluated by two linguistic experts

(student assistants) per sample according to three different aspects: parsability of a sentence, number, and accuracy of roots, and parsing accuracy itself.

4.3.4.1 Parsability

We evaluated 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 accepted title-like noun phrases (Example (2)) as well as dates (Example (3)), but we excluded sentences in languages other than English or German (Example (4)) and fragments without grammatical, linguistically parsable structure such as equations (see Example (5)), as well as accumulations of abbreviations (Example (6)).

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

(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 (Table 4.9) show that for both corpora (RSC, English; DTAW, German) the selection of "good sentences" (**GS**) was 100% successful, i.e., all of the retained sentences are parsable. The numbers for parsability of the "bad sentences" (**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 found the opposite trend: our preprocessing excluded more actually parsable sentences from newer data than in the older data. This is due to a much higher number of foreign language 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 that are still syntactically interpretable. All resulting parsability values for "bad sentences" are significantly below the values obtained for "good sentences".

4.3.4.2 Roots

We furthermore evaluated the number and accuracy of roots per sentence. A well-parsed sentence should only have one root. We checked how many roots were assigned to one sentence and evaluated if the assignment was correct. We found that for English, UDpipe consistently assigned exactly one root to each of the **GS** while assigning more than one root to the **BS** (Table 4.10). Also, the precission was significantly

	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 4.9: Evaluation of parsability of a sentence:
Statistics for RSC: $t = 5.55$, $df = 8$, $p < 0.0005$.
Statistics for DTAW: $t = 4.12$, $df = 8$, $p = 0.0033$

higher for the **GS** than for the **BS** (see Table 4.11). For German, root detection did not seem to work very well, neither for the **GS** nor for **BS** (see Table 4.10), which is also reflected in the fact that average numbers of roots per sentence in **GS** and **BS** do not vary significantly. This also shows that the processing does not improve the one-root-per-sentence-only processing of the German parser. The detection of several roots per sentence in German therefore seems instead to be due to parser-internal issues. However, the accuracy of root detection (Table 4.11) is significantly better for the **GS** than for the **BS**.

	RSC		DTAW	
Period	GS	BS	GS	BS
1650–1699	1	1.30	1.35	1.45
1700–1749	1	1.30	2.50	1.45
1750–1799	1	1.35	1.40	1.65
1800–1849	1	1.05	1.20	1.35
1850–1899	1	1.05	1.25	1.50
mean	**1**	**1.21**	**1.54**	**1.48**

Table 4.10: Number of roots per sentence:
Statistics for RSC: $t = 3.1840$, $df = 8$, $p = 0.0129$.
Statistics for DTAW: $t = 0.2424$, $df = 8$, $p = 0.8145$.

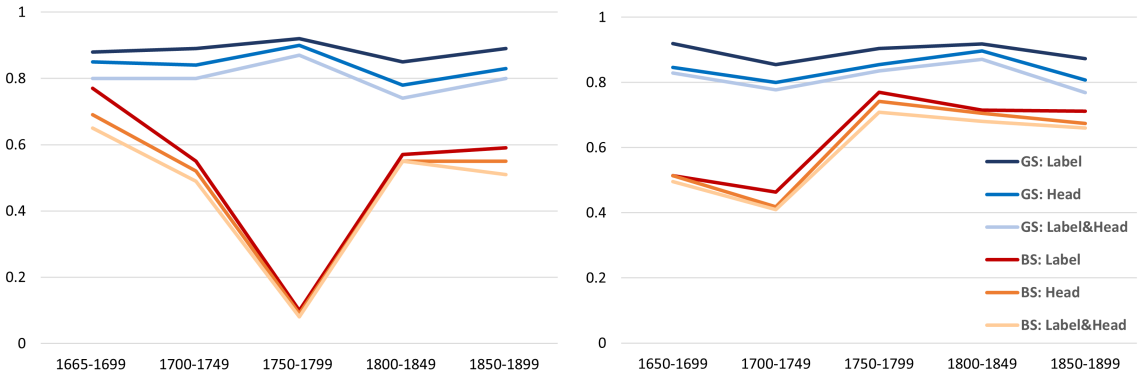
4.3.4.3 UD-annotation

Following the example of SpaCy's accuracy evaluation[9], we evaluated the correctness of the assigned UD-label (Label) per token (cf. DEP/LAS in Spacy's evaluation scheme), the correctness of the syntactic head (Head) of each token (cf. DEP/UAS

[9]Documented at SpaCy.io by Honnibal & Ines (2022).

	RSC		DTAW	
Period	GS	BS	GS	BS
1650–1699	0.70	0.25	0.59	0.38
1700–1749	0.80	0.15	0.38	0.28
1750–1799	0.80	0.15	0.68	0.39
1800–1849	0.65	0.35	0.88	0.48
1850–1899	0.85	0.15	0.64	0.47
mean	**0.76**	**0.21**	**0.63**	**0.40**

Table 4.11: Precision of detected roots:
Statistics for RSC: $t = 10.1263$, $df = 8$, $p < 0.0001$.
Statistics for DTAW: $t = 2.66$, $df = 8$, $p = 0.0290$.



(a) RSC **(b)** DTAW

Figure 4.8: Accuracy of UD Label and Head in (a) RSC and (b) DTAW by 50-year periods.

in Spacy's evaluation scheme), and correctness of both labels (Label and Head) per token. Accuracy was calculated as the number of correctly annotated tokens over the whole number of tokens in a time period. We conducted evaluations for **GS** as well as **BS**. The parse of a non-parsable sentence was regarded as entirely incorrect, since for such a sentence no actual correct parse exists. Figure 4.8 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 preprocessing improves parsing accuracy significantly.
For English (Table 4.12), the 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 did not find an accuracy improvement over time; in fact, t-tests for all time periods compared to each other show no significant difference in the accuracy values encountered for each period. Looking at the English **BS**, we see that parsing

quality drops towards the end of the 18thc. and increases afterward (Figure 4.8). 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 (e.g. *n-1 X 1/~~1.*), reducing parsability.

	Label		**Head**		**Label&Head**	
Period	GS	BS	GS	BS	GS	BS
1665–1699	0.88	0.77	0.85	0.69	0.80	0.65
1700–1749	0.89	0.55	0.84	0.52	0.80	0.49
1750–1799	0.92	0.10	0.90	0.09	0.87	0.08
1800–1849	0.85	0.57	0.78	0.55	0.74	0.55
1850–1899	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 4.12: Evaluation of parses of good sentences (GS) vs. bad sentences (BS) in the RSC: 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 4.13: Evaluation of parses of good sentences (GS) vs. bad sentences (BS) in the DTAW: correct UD-label, correct recognition of syntactic head, correct UD-label and head.

For German GS, we found slightly higher accuracy for Label and Head than for the English data (see Table 4.13) with values between 80 and 90%. Just 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 preprocessing 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 preprocessing steps help improve parsing quality significantly on all three levels: parsability, root accuracy, and UD-annotation (Label and Head detection). For English, our preprocessing also contributes significantly to preventing parses from containing more than one root.

4.3.5 Final annotation

After UD-parsing of the corpora, we calculated the dependency length (DL) of each token as described in Section 5.2.2 and annotated it as a positional attribute to each token as an additional column in the CONLLU format. We furthermore calculated and annotated the sum of dependencies (SDL) as the sum of all DLs in one sentence, the sentence length (SL) in tokens (excluding punctuation), and the average DL (ADL) in each sentence as a structural attribute of each sentence in the corpus. We furthermore calculated a new surprisal on the parsed versions of the corpora resulting in the following complete list of linguistic annotations on our UD-parsed corpora (Table 4.14).

	attribute	description
positional attributes	word	-
	lemma	-
	upos	Part-of-Speech using Universal Depencies
	pos	Part-of-Speech using PennTreebank/STTS tagset
	ufeat	Universal Features (morphological annotation)
	parent	the parent of a token in the dependency tree
	urel	Universal Dependency Relation
	DL	Dependency length
	srp	Surprisal
	srp_avg	Average surprisal
structural attributes	SDL	Sum of dependency lengths
	SL	Sentence length
	ADL	Average dependency length

Table 4.14: Annotation of parsed corpora.

Chapter 5

Complexity Measures

In this chapter, we explain the different measures to trace complexity affecting processing effort (as introduced in Section 2.1.3) used in this thesis. In Section 5.1, we start by presenting the measures associated with the two indicators of *lexico-grammatical complexity*, i.e. *entropy* (Section 5.1.1) indicating the *paradigmatic richness* of the relativizer paradigm and *surprisal* (Section 5.1.2) accounting for the *syntagmatic predictability* of a relativizer in its syntagmatic context and thus approximating the expectation-based processing effort when encountering a relativizer at a given choice point. In Section 5.2, we will present the complexity measures associated with the three different indicators of *syntactic complexity*, i.e. *intricacy* (Section 5.2.1) as indicated by the frequency of RCs in a corpus as well as within single sentences, *locality* (Section 5.2.2) as measured by dependency length (DL), and *accessibility* (Section 5.2.3) as measured by the frequency of different RC extraction types accounting for expectation-based processing effort.

5.1 Measuring lexico-grammatical complexity

As stated in Section 2.1.3.1, we use two indicators of lexico-grammatical complexity: *syntagmatic predictability* and *paradigmatic richness*. In the present chapter, we will present how the corresponding measures to determine these complexity indicators are calculated. We start by presenting the first complexity measure, entropy (H), which we calculate over the paradigm of relativizers as a measure of *paradigmatic richness*. Next, we explain how the 3-gram surprisal of the relativizers *which/welch*, *that/d* and a group of pronominal adverbs per 50-year segment is calculated to estimate *syntagmatic predictability*.

5.1.1 Paradigmatic richness – Entropy

We analyze register-specific preferences for relativizers as an indicator of grammatical complexity, which we assume to decrease over time in the scientific literature for the sake of lower processing effort due to uncertainty about the upcoming word (H 1.1, Section 3.1.1.1). For this, we use entropy (H) as an indicator of uncertainty about a set of choices at a given point.

Entropy can be calculated in different ways. One major distinction between entropy calculations is non-conditional entropy and conditional entropy. As the name suggests, the latter is calculated based on conditional probabilities, i.e. taking into account the preceding context (i.e., an n-gram probabilistic model: see e.g. Genzel & Charniak, 2002) of a word. The former is calculated on the general probabilistic distribution of a certain group of words. In the present study, we calculate the ***non-conditional entropy*** of the paradigm of relativizers based on the general probabilistic distributions of each relativizer type as it occurs in a corpus.

In the present thesis, entropy (H) represents the expected amount of information in a relativizer paradigm: Entropy depends on the number of members in the paradigm and on the probability distributions of the members. Consequently, the fewer members a paradigm has and the more skewed their probabilities are (i.e. favoring one option), the lower the entropy of the paradigm. Formally, entropy is calculated as follows:

$$H = -\sum_{i=1}^{m} p_i \log_2 p_i \tag{5.1}$$

Consider the simplified example of entropy over the hypothetical paradigm consisting of only *that* and *which*. To calculate the paradigm's entropy, we need to know the raw frequencies $f(W_i)$ of the two relativizers in a corpus, i.e., let the frequency of *which* be $f(which) = 42,000$ and the frequency of *that* be $f(that) = 167,000$. The probability of the individual members of the paradigm p_i is calculated as the frequency of the member $f(W_i)$ divided by the whole number of members within the paradigm $f(W)$. We can now calculate entropy by inserting the frequencies into the formula:

$$\begin{aligned} H &= -\sum_{i=1}^{m} \frac{f(W_i)}{f(W)} \log_2 \frac{f(W_i)}{f(W)} \\ &= -\left[\left(\frac{42,000}{209,000} \log_2 \frac{42,000}{209,000}\right) + \left(\frac{167,000}{209,000} \log_2 \frac{167,000}{209,000}\right)\right] \\ &= -\left[(0.2 \times \log_2 0.2) + (0.8 \times \log_2 0.8)\right] \\ &= -[-0.464 - 0.251] \\ &= 0.715 \end{aligned} \tag{5.2}$$

If the distribution was more skewed towards one of the relativizers, i.e. $f(which) = 2,000$ and $f(that) = 207,000$, the entropy would be lower:

$$
\begin{aligned}
H &= -\sum_{i=1}^{m} \frac{f(W_i)}{f(W)} \log_2 \frac{f(W_i)}{f(W)} \\
&= -\left[\left(\frac{2,000}{209,000} \log_2 \frac{2,000}{209,000}\right) + \left(\frac{207,000}{209,000} \log_2 \frac{207,000}{209,000}\right)\right] \\
&= -\left[(0.01 \times \log_2 0.01) + (0.99 \times \log_2 0.99)\right] \\
&= -[-0.07 - 0.01] \\
&= 0.08
\end{aligned} \tag{5.3}
$$

If the distribution was 50:50 for both relativizers, the entropy would be highest, namely 1:

$$
\begin{aligned}
H &= -\sum_{i=1}^{m} \frac{f(W_i)}{f(W)} \log_2 \frac{f(W_i)}{f(W)} \\
&= -\left[\left(\frac{104,500}{209,000} \log_2 \frac{104,500}{209,000}\right) + \left(\frac{104,500}{209,000} \log_2 \frac{104,500}{209,000}\right)\right] \\
&= -\left[(0.5 \times \log_2 0.5) + (0.5 \times \log_2 0.5)\right] \\
&= -[-0.5 - 0.5] \\
&= 1
\end{aligned} \tag{5.4}
$$

In Chapter 6, we investigate the distributions of the different relativizers in scientific and mixed genre texts of English and German. We apply entropy to both relativizer paradigms 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 decreasing processing effort (compare H1.1, Section 3.1.1.1).

5.1.2 Syntagmatic predictability – Surprisal

Surprisal (Shannon, 1948; Hale, 2001; Levy, 2008) describes the information content of a linguistic unit (e.g. word) indicating "the extent to which the word came unexpected to the reader or listener" (Frank, 2013). Surprisal has been shown to correlate with the cognitive processing effort of a linguistic unit (e.g. Frank & Frank, 2009). This correlation has often been proven by measuring processing effort in terms of reading times (e.g. Demberg & Keller, 2008). Formally, surprisal is an information-theoretic measure indicating the number of bits needed to encode a message. For linguistic research, this number of bits transmitted in a particular linguistic unit (here words[1]) is calculated by the unit's probability given its preceding syntagmatic context. In our case, we calculate surprisal based on the conditional negative *log* probabilities from

[1]Our surprisal calculation also treats punctuation marks as words.

a 4-gram language model, i.e. the negative *log* probability of a word given its three preceding words:

$$S = -\log_2 p(w_n \mid w_{n-3}w_{n-2}w_{n-1}) \tag{5.5}$$

The higher the probability of a given word in a particular context, the fewer bits are needed to encode it and the less surprising is its occurrence in this particular context. A word with a very low probability in a certain context requires more bits to be encoded and is thus more surprising. Consider the following examples of two possible contexts preceding a relative clause for illustration:

(1) a. *by means of* ***which***
b. *the integrity of* ***which***

Comparing Example (1-a) to (1-b), intuition tells us that *which* is much more predictable given *by means of* than *the integrity of*. This is because, in English, (1-a) occurs more often than (1-b), giving the former a higher probability than the latter. Let us assume (1-a) occurs 20 times and the context *by means of* occurs 50 times, whereas (1-b) only occurs once while the context *the integrity of* occurs 20 times. The *conditional* probability p of (1-a) would be $20/50 = 0.4$ and p of (1-b) would be $1/20 = 0.05$. Thus the surprisal of (1-a) would be calculated as follows:

$$\begin{aligned} S &= -\log_2 p(\textbf{\textit{which}}|\textit{by means of}) \\ &= -\log_2(0.4) \\ &= -(-1.322) \\ &= 1.322 \text{ bits} \end{aligned} \tag{5.6}$$

Thus, the number of bits needed to encode *which* in (1-a) would be 1.3 bits. Similarly, the surprisal of (1-b) would be calculated as follows:

$$\begin{aligned} S &= -\log_2 p(\textbf{\textit{which}}|\textit{the integrity of}) \\ &= -\log_2(0.05) \\ &= -(-4.322) \\ &= 4.322 \text{ bits} \end{aligned} \tag{5.7}$$

So, fewer bits are needed to encode a word when the number of occurrences of the word in the same context is higher (compare 1.3 bits vs. almost 4.3 bits) than when a word rarely occurs in a certain context. In terms of information content, a word that occurs less often in a certain context is more surprising and thus carries a higher amount of information, in terms of bits of information. In our analyses, we are interested in the distributions of the surprisal values of three groups of relativizers (*which/welch.**, *that/d.** and *pronominal adverbs*) across time, i.e. we ask how the surprisal of a word in context changes over time.

The comparison between obtained surprisal values is, however, not trivial, since our language model is trained on 50-year slices of a corpus and tested on each text belonging to the 50-year slice. In this sense, technically speaking, the surprisal values that we obtain belong to five different corpora, i.e. the five 50-year periods 1650–1699, 1700—1749, 1750–1799, 1800–1849, and 1850–1899. Now, surprisal can be used for comparing the predictability of different sequences of words within the same modeling space, i.e. the corpus that it was trained on, but it is not directly comparable across different corpora for several reasons. On the one hand, different corpora may have different vocabularies (i.e. the number of different types in a corpus) and word frequencies, which can affect the predictability the language model assigns to each sequence. For example, if one corpus has a higher frequency of rare words than another corpus, then the surprisal values for those words will be higher. Secondly, the size and composition of the corpus can also affect the predictability of the language model. A larger corpus with more diverse texts may result in lower surprisal values compared to a smaller corpus with limited text genres.

For this reason, we only compare the surprisal values of the different relativizers within a period. We do so by inspecting box plots illustrating the distributions of surprisal values in a 50-year period. Since box plots show the mean and median of the distribution as well as the interquartile range (IQR), we can compare the distributions between two items in one period, e.g. how surprisal of *which* is distributed compared to the surprisal values encountered for *that*. To make statements about surprisal differences between time periods, we calculate the differences between the median surprisal of at least two different items in one 50-year period, e.g. the median surprisal of the relativizer *which* and that of the relativizer *that* in each period. Let us assume the median surprisal value of *which* between 1650 and 1700 is 3.2 and that of *that* is 2.7, resulting in a difference of 0.5. If in the period between 1850 and 1899, the surprisal of *which* is 2.7 and that of *that* is 2.7, resulting in a difference of 0, we can claim that despite the expected differences in vocabulary size, the median surprisal value of *that* has stayed stable while the median surprisal of *which* has declined, leading to a decrease of the intra-periodical surprisal difference. To statistically evaluate the differences between two items in period A and period B, we conduct one-sided t-tests to demonstrate that the difference between intra-periodical differences is significant.

5.2 Measuring syntactic complexity

As stated in Section 2.1.3.2, we look at three indicators of syntactic complexity: *intricacy*, *locality*, and *accessibility*. In the present chapter, we will present how the corresponding measures to determine these complexity indicators are calculated. We start by presenting *intricacy* calculated as the relative frequency of RCs per 50-year period and as the average number of RCs per sentence (embeddedness). Next, we explain how dependency length (DL) as a measure of *locality* was calculated,

and finally we describe how the distributions of different RC types as a measure of *accessibility* are derived.

5.2.1 Intricacy – Relative clause frequency and embeddedness

To analyze RC frequency and embeddedness, we use the parsed corpus versions described in Section 4.3.2. For the calculation of the relative frequencies of RCs per 50-year period, we first extract all RCs per period by searching the corpora for the Universal Dependencies relation `acl:relcl`. We calculate relative frequencies normalized per 1000 sentences and for comparison normalized per 1 million words. To calculate embeddedness, we extract each sentence including at least one RC, and calculate the average number of RCs per sentence including RCs.

5.2.2 Locality – Dependency length

As introduced in Section 2.1.3.2, dependency length (DL) is positively correlated with memory-based processing effort and is thus a good measure for determining processing-related syntactic complexity. We have shown in Section 4.3 how the corpora were annotated with Universal Dependencies (UD). In the present section, we present how DL and derived values (summed DL, i.e. SDL and average DL, i.e. ADL) are calculated. The calculation of DL strongly depends on the underlying syntactic framework. While dependency-grammar-based approaches define DL in terms of the number of intervening words between the syntactic head and its dependent (Heringer et al., 1980; Hudson, 1995; Wasow, 2002), PSG-based approaches take the number of intervening discourse referents (Gibson, 1998) as a metric of distance. There are even approaches measuring the distance in terms of intervening syllables or the number of lexical stresses (Anttila et al., 2010). Subsuming these measures as measures of weight, Grafmiller & Shih (2011) point to a strong correlation among these measures.

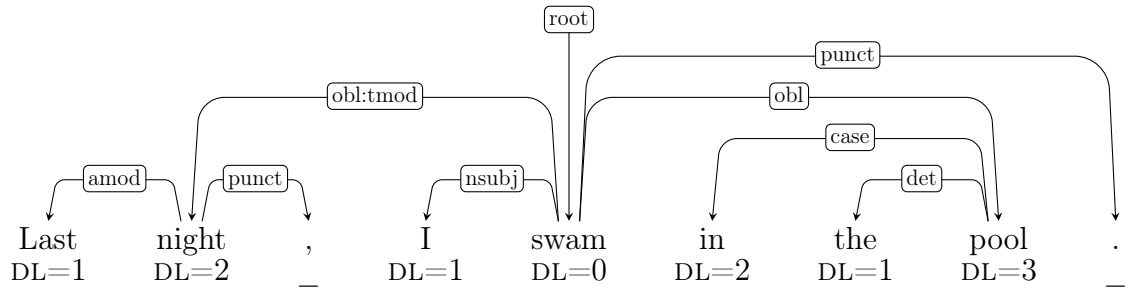


Figure 5.1: Graphic visualization of a simple sentence in the Universal Dependencies framework. The edges represent a dependency relation pointing from head to dependent; the numbers denote the dependency length (DL) between tokens.

A

Since our data are annotated with Universal Dependencies, our calculation of DL is based on Dependency Grammar, defining DL as the number of intervening words between a syntactic head and its dependent. The metric draws inspiration from the research conducted by Futrell et al. (2015) and Gibson et al. (2019) and expands upon earlier work by Liu (2008). The measure proposed by Liu (2008) differs slightly in how the average is calculated. In our calculation of DL, we exclude punctuation altogether since punctuation has changed widely over the past centuries and comparability of DL over time would be problematic. For illustration, consider the example in Figure 5.1. The DLs between two tokens are marked by the numerals below. DL in tokens between head and dependent is calculated as follows: for any sentence s of length n, we can calculate the distances (DL) for all tokens t_1 to t_n by subtracting a token's position (t.id) from its head (t.hd), and then subtracting any intervening punctuation between t.hd and t.id. This can be expressed mathematically as:

$$\text{DL}(t_i) = |\text{t.id} - \text{t.hd}| - \sum_{j=\text{t.hd}+1}^{\text{t.id}-1} \text{punctuation}(j) \tag{5.8}$$

In this formula, $\text{DL}(t_i)$ represents the distance for the ith token t_i. The absolute difference between the token's position (t.id) and its head (t.hd) is calculated using $|\text{t.id} - \text{t.hd}|$. The sum over j computes the total amount of intervening punctuation between the head and the position of the token, which is then subtracted from the absolute difference to obtain the final distance metric. The subscript and superscript in the sum notation specify the range of values that the index j can take. Thus, the DL from the head *swam* to the temporal modifier *night* is 2 and the DL to the oblique nominal *pool* is 3. Next, to calculate ADL of a sentence, we need to calculate the sentence length (SL). Also in SL, punctuation is excluded:

$$\text{SL}(s) = \sum_{i=1}^{n} \text{token}(i, s) - \sum_{j=1}^{m} \text{punctuation}(j, s) \tag{5.9}$$

According to this definition, the SL of the example sentence is 7 and not 9 tokens. To measure the syntactic complexity of a sentence, we use the *average dependency length* (ADL) of a sentence calculated as the sum of all DL per sentence (SDL) divided by the SL. ADL per sentence is thus calculated as follows:

$$\text{ADL} = \frac{\sum_{i=1}^{n} \text{DL}(i)}{\text{SL}(s)} \tag{5.10}$$

For the present example, the SDL amounts to $1+2+1+2+1+3 = 10$ and $\text{SL} = 7$; thus $\text{ADL} = 10/7$. The ADL can be interpreted as a proxy of the (cumulative) processing difficulty of an entire sentence. However, ADL is, of course, a function of sentence length, i.e. only long sentences allow very long DLs to be built. Thus, the ADL has to be normalized by SL. When looking at syntactic complexity diachronically, we take

the average of the ADL per 50-year period. In this case, it is especially important to interpret the ADL depending on the underlying SL, since over time the distribution of SLs changes, i.e. over time, there are more short sentences and fewer long sentences. When analyzing specific dependency relations such as RCs (`acl:relcl`), we either analyze their ADL normalized per SL, or we inspect it for one particular SL.

5.2.3 Accessibility – Relative clause type

To measure the *accessibility* of an RC, i.e. expectation-based syntactic complexity, we look at the diachronic distribution of the possible extraction types of RCs, i.e. the NP positions they can relativize (as displayed on the Accessibility Hierarchy (AH), Keenan & Comrie, 1977). According to the AH, the processing is harder the further down the hierarchy the RC is extracted from. The type of RC is indicated as the dependency relation annotated on the relativizer, and thus it can be determined by extracting the relativizers and their UD-relations as in the following examples (Figure 5.2):

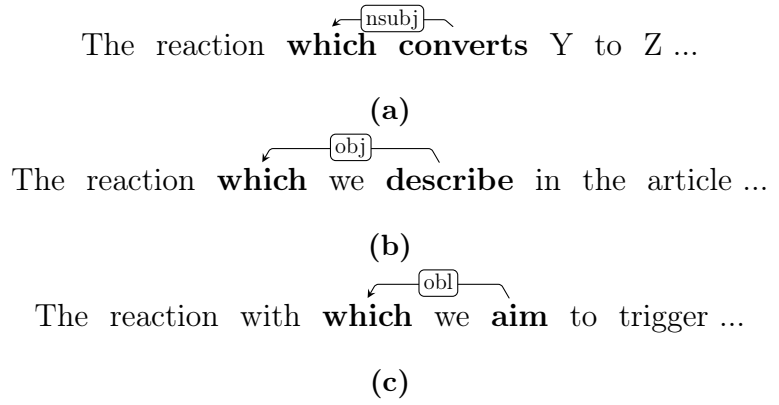


Figure 5.2: RC types: (a) Subject RC. (b) Direct Object RC. (c) Oblique RC.

In addition to analyzing the temporal distributions of the various types of relative clauses, we developed a metric called the "accessibility score" (*a-score*) to quantify the overall accessibility of RCs within a specific 50-year period. To obtain the *a-score*, we assign a value (v) to each of the UD-relations. Note that subject RCs can be divided into active subject RCs (`nsubj`) and passive subject RCs (`nsubj:pass`). For calculating the *a-score*, they are both assigned the same value. Also, note that the UD-annotations for English and German differ slightly: In English, direct objects are annotated as `obj`, and indirect objects are subsumed to `obl`. In German, the UD-tagset differentiates between direct objects (`obj`) and indirect objects (`iobj`). So in German, v can take the following values: subject RC = 1, direct object RC = 2, indirect object RC = 3, oblique object RC = 4. In English, v is assigned as follows: subject RC = 1, object RC = 2, oblique object RC = 3. We then multiply the frequencies of the UD-relations of a relativizer (f_i) with their corresponding factors (v)

and sum the resulting products and divide them by the total number (n) of UD-relations.

$$a\text{-}score = \frac{1}{n} \sum_{i=1}^{n} f_i \cdot v_i \tag{5.11}$$

In Equation 5.11, f_i represents the frequency of the ith UD-relation of a relativizer, and v_i represents the corresponding factor for that dependency relation. The summation iterates over all n UD-relations of all relativizers, and the entire summation is divided by the total number of UD-relations n of all relativizers. Thus, the *a-score* in German can take values between 1 and 4, while in English the *a-score* is between 1 and 3.

While the *a-score* is an aggregate measure of accessibility showing on average how accessible RCs are in a 50-year period, it does not show which of the RC types is the most influential in this development. Thus, as noted above, we also look at the distributions of the different RC types, assuming that more accessible RC types (i.e. subject RCs) become more frequent and less accessible ones decrease in frequency.

Part III

Corpus Studies: Lexico-grammatical Complexity

Chapter 6

Paradigmatic Richness

In this first chapter of our corpus analyses, we focus on the development of the set of relativizers in English and German between 1650 and 1900. As stated in Hypothesis H1.1 (Chapter 3), we assume that in scientific language the paradigm of relativizers will be adapted to register-specific needs by converging on specific, well-suited options to introduce RCs. This convergence leads to lower lexico-grammatical complexity in terms of *paradigmatic richness* and thus makes scientific writing more efficient by avoiding uncertainty about which relativizer will be chosen. We calculate the degree of this uncertainty using entropy (cf. Section 5.1.1) and show that, while some relativizers will be preferred over time, other less preferred ones will be abandoned[1].

We start by determining the respective members of the relativizer paradigms in English and German in Section 6.1. We will test our assumptions on the reduction of paradigmatic richness by conducting a macro-analysis looking at the entropy (Section 6.2) of the relativizer paradigm per 50-year period[2]. We assume that entropy will decrease due to the higher predictability of one preferred option alongside the lower predictability of other, increasingly dispreferred options leading to a stronger skew in probabilities across the paradigm. Next, we will analyze the obtained entropy values by analyzing the frequency distributions of the different relativizers (Section 6.3). Specifically, we are interested in which options scientific language converges on, and which options are abandoned. We summarize our findings for both languages in Section 6.4.

6.1 Determination of the paradigm

In this first analysis, we inspect lexico-grammatical complexity through the lens of paradigmatic richness, focusing on relativizers in English and German. Obviously,

[1]Part of the study has previously been published in (Krielke, 2021).

[2]Periods cover 50 years. For space economy in the figures they are referred to as follows: 1650 = 1650–1699, 1700 = 1700–1749, 1750 = 1750–1799, 1800 = 1800–1849, 1850 = 1850–1899.

linguistic change also affects paradigms. Thus, in order to determine the size of a paradigm in each 50-year period, we first need to determine which members belong to it at different stages of time and in different meta-registers, i.e. scientific and general language.

The term relativizer can refer to relative pronouns such as *which/welch(e/er/es)*, *that/d(er/ie/as)*[3], *who(m/se)/wer*, and *what/was*, as well as to relative adverbs such as *where/wo*, *why/warum*, *when/wann* and *how/wie*, and *w(h)*-pronominal adverbs (PAs) such as *whereby/wobei*, *whereof/wovon*, etc. In the present thesis, we are only interested in relativizers that refer to non-human nominal antecedents. Since for German (apart from *wer*), there is no designated relativizer to refer to human antecedents, we are not able to control for humanness as precisely as for English. However, a random sample of 50 instances of relativizers from the mentioned group shows that in our scientific German texts, only 8% of antecedents are human, while in general German texts, the proportion is 10%. We exclude relative adverbs from our analyses due to their strong ambiguity and often erroneous annotation (interrogative pronoun vs. relative adverb). We thus concentrate on the group of relativizers consisting of *which/welch(e/er/es)*, *that/d(er/ie/as)*, *what/was* and PAs, i.e. compound words consisting of *where/wo(r)* + preposition.

To grasp the full historical extent of the paradigms apart from those relativizers still in use in present-day English and German, we first determine the existing members of the group of PAs. For English, we extract all words beginning with *where-* from the general and the scientific English corpus and sort out all words not representing PAs. For German, we extract all words beginning with *wo(r)-* and part-of-speech (POS) tagged as PRELS/PRELAT, resulting in the lists provided in Table 6.1. To complete the list of the paradigms we focus on in this thesis, we also extract the standard relativizers referring to non-human antecedents (*which*, *that*, *what* and *welch.**, *d.** and *was*). The resulting word list is used for all the following analyses in the present Section. In the analysis, we compare the trends of entropy in scientific and general language to find out whether scientific language shows a register-specific trend of reducing the paradigmatic richness of relativizers as compared to general language.

[3]and all respective inflections of the German relativizers. **In the thesis, all inflectional forms of *welch-* and *d-* will be referred to as *welch.** and *d.*.**

	English	German
Standard relativizers	which	welch.*
	that	d.*
	what	was
Pronominal adverbs (PAs)	whereabouts	worüber
		woher
	whereat	woran
	whereby	wobei
	wherefore	wofür
	wherefrom	wovon
	wherein	worin
	whereof	woraus
	whereon	worauf
	whereout	wodurch
	whereto	wohin
	whereupon	woraufhin
	wherewith(al)	womit
	whereinto	wohinein
		wogegen
		woherum
	whereunto	wohinauf
		wohinaus
		wohin
		wonach
		woneben
		worein
		worum
		worunter
		woselbst
		wovor
		wozu
		wozwischen

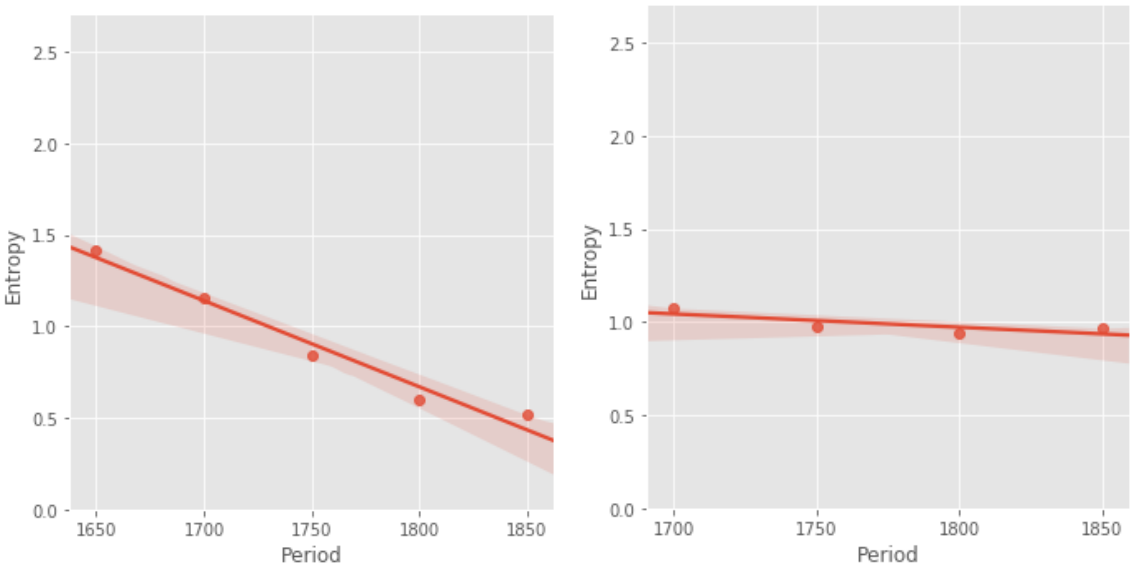
Table 6.1: Members of the relativizer paradigms in English and German.

6.2 Entropy

To calculate the entropy over the relativizer paradigms in each corpus, we extract all relativizers from the corpus as well as all words in the corpus. We then calculate the probability of each relativizer to occur in each time period amongst all other words in that same time period. Inserting the resulting probabilities in the entropy

formula (Equation 5.1, Section 5.1.1) gives us the entropy (uncertainty) that a reader encounters at the choice point of a relativizer in a specific 50-year period.

6.2.1 English



(a) RSC **(b)** CLMET

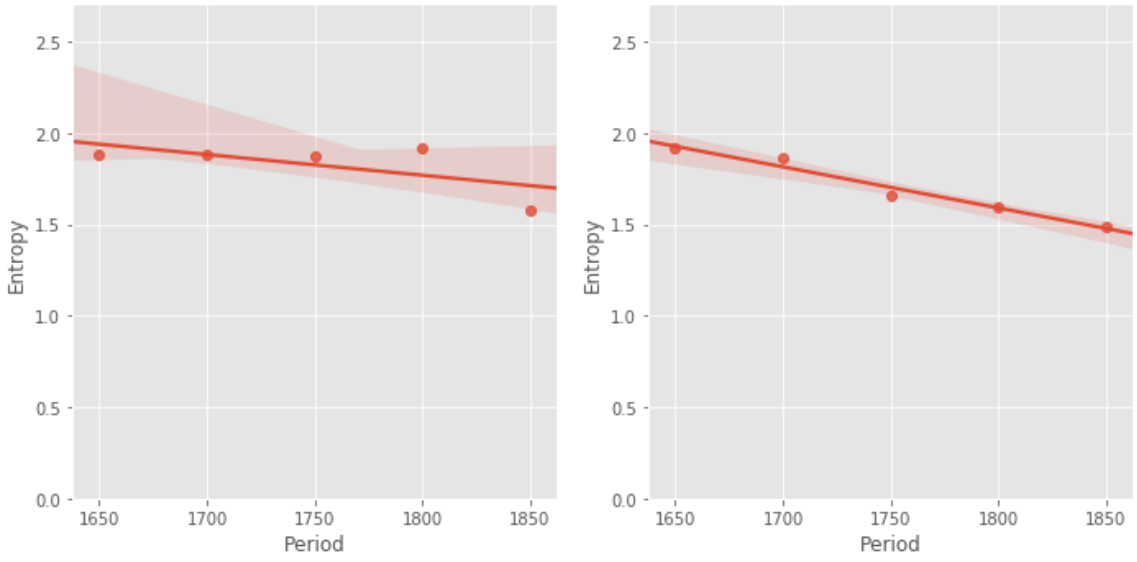
Figure 6.1: Development of entropy of the relativizer paradigm in (a) scientific (RSC) and (b) general (CLMET) English. Each data point represents the entropy value of the paradigm per 50-year period.

For English, we thus extract all relativizers belonging to the English relativizer paradigm from the two English corpora (RSC for scientific English and CLMET for general English) and calculate the entropy for both corpora per 50-year period. According to our H1.1 (Section 3.1.1.1), we expect to find differences in the development of scientific and general English, where scientific English should show a straight reduction in entropy of the relativizer paradigm while paradigmatic richness in general English is not expected to change much. Starting with scientific English, (Figure 6.1a), we see a striking reduction of entropy (almost one bit) over the observed time span. General English (Figure 6.1b), on the other hand, shows relatively stable entropy values. Both trends confirm our hypothesis H1.1 that during register formation in scientific English, the uncertainty about the choice of a relativizer decreases leading to a reduction in processing load associated with entropy (Milin et al., 2009). The comparison with general English (exhibiting comparatively consistent levels of entropy over time) shows that this reduction in entropy is specific to the scientific meta-register. This observation is in line with our hypothesis H1 (Section 3.1) that register formation should lead to diverging developments in scientific vs. general language due to the continuous specialization of the emerging scientific meta-register. We assume

that scientific language is subject to stronger communicative pressures than general language in terms of lexico-semantic expansion, necessitating compensation on the lexico-grammatical level.

6.2.2 German

As done for English, we first extract all relativizers belonging to the German relativizer paradigm from the two German corpora (DTAW for scientific German and DTAG for general German) and calculate the entropy for both corpora per 50-year period. According to our H1, we expect to find a distinct development of scientific and general German, with scientific German first becoming more complex in terms of paradigmatic richness as indicated by an increase in entropy towards 1800, and exhibiting a decrease thereof afterward. We expect this time-shifted development on the grounds of previous work suggesting that scientific vernacular German started to develop as an independent register much later than English, due mainly to the fact that most scientific communication in the German-speaking area until the beginning of the 19th c. was written in Latin.



(a) DTAW **(b)** DTAG

Figure 6.2: Development of entropy of the relativizer paradigm in (a) scientific (DTAW) and (b) general German (DTAG). Each data point represents the entropy value of the paradigm per 50-year period.

Both scientific German (DTAW, Figure 6.2a) and general German (DTAG, Figure 6.2b) show overall higher entropy values compared to English. A reason for this may be the generally higher number and a more even probability distribution of relativizers available in German compared to English. Scientific German shows remarkably stable entropy values until the period of 1750 (at approx. 1.9 bits). In

the period of 1800, entropy even increases slightly, almost reaching 2 bits, and it then falls remarkably in the period of 1850 to slightly above 1.6 bits. This trend is especially interesting when compared to the entropy trend in scientific English showing a strong and linear decrease. The initial slight increase in entropy in scientific German points to a turn from previously more even probability distribution over the different relativizers towards higher probabilities of some more probable options and other less probable ones over time. The fact that entropy in scientific German ultimately declines confirms our hypothesis H1.1, which states that paradigmatic richness decreases over time in scientific German. The time-shifted drop in entropy after 1850 confirms our hypothesis H2 (Section 3.2) stating that the turn towards lower complexity appears later in German than in English. In contrast to scientific German, in general German, the entropy of the relativizer paradigm steadily decreases after 1650, so that in 1850, entropy in scientific and general German is almost equally low. Comparing the entropy trajectories in scientific and general German, the results are rather surprising against the backdrop of our hypothesis H1 that due to register formation, scientific language should evolve towards lower complexity than general language. Our results suggest that this decreasing trend in paradigmatic richness is actually initialized in general German, whereas scientific German instead seems to follow this trend.

6.3 Frequency distribution

We now move on to analyzing the resulting entropy values by looking at the frequencies of the members of the paradigm. The underlying frequency and distributional configurations of the relativizer paradigm in the corpora show which relativizers become preferred options and which ones become dispreferred.

6.3.1 English

In Section 6.2.1, we have discovered that in line with our expectations, entropy drops in scientific English while staying stable in general English. This reduction in entropy is necessarily the result of changes in the relativizer paradigm in the scientific metaregister. Let us therefore now take a look at the development of the distributions and relative frequencies of the different relativizers over time.

Figures 6.3 and 6.4 show the percentage distributions of the relativizers per 50-year period. Scientific and general English differ substantially regarding their relativizer distributions. For scientific English (Figure 6.3), we find that the development of relativizers over the two centuries of the Late Modern Period shows a remarkable trend towards choice reduction. The percentage distributions of the different members of the relativizer paradigm change strongly over time. The first period (1650: 1665–1699) starts out with a great variety of available relativizers including a large group of PAs. The relativizer *which* is proportionally the strongest, followed by *that*. Over time,

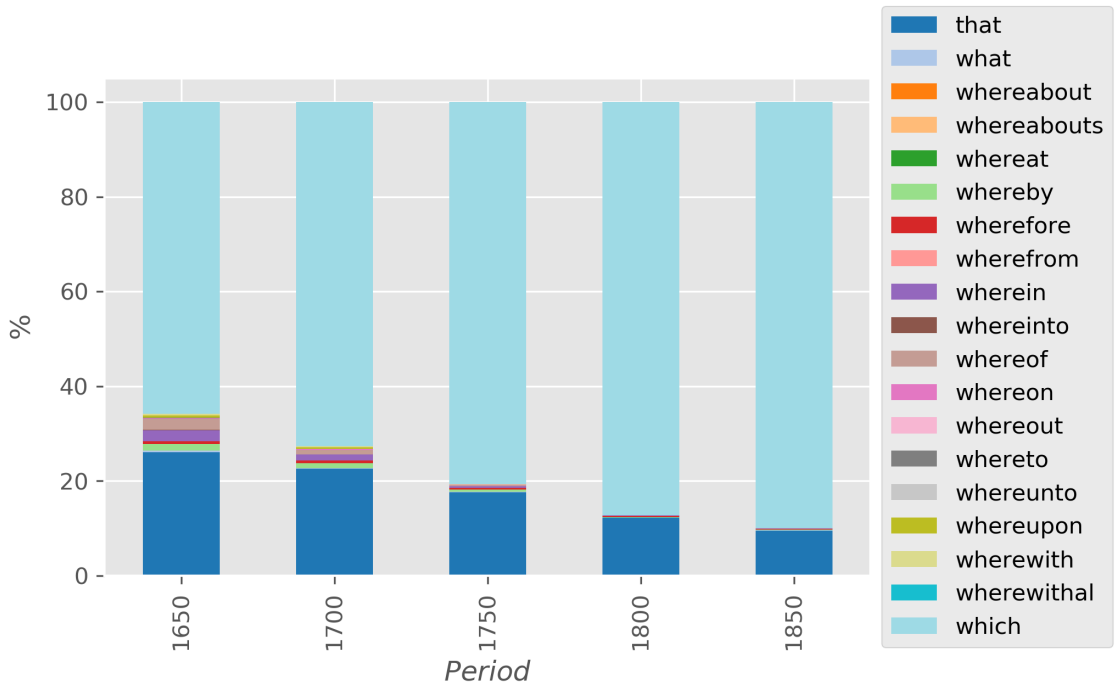


Figure 6.3: Percentage distribution of relativizers in scientific English (RSC).

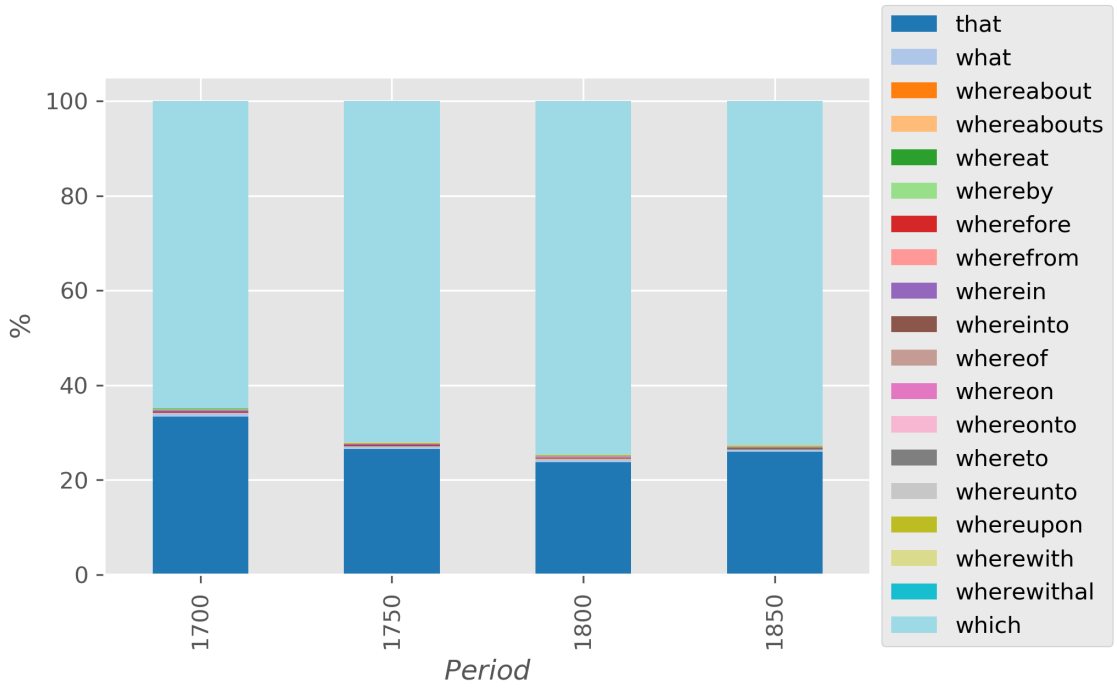


Figure 6.4: Percentage distribution of relativizers in general English (CLMET).

which further increases proportionally, and nearly pushes out all other alternatives to under 10% in the period of 1850, the most notable demise of alternatives being that of the PAs.

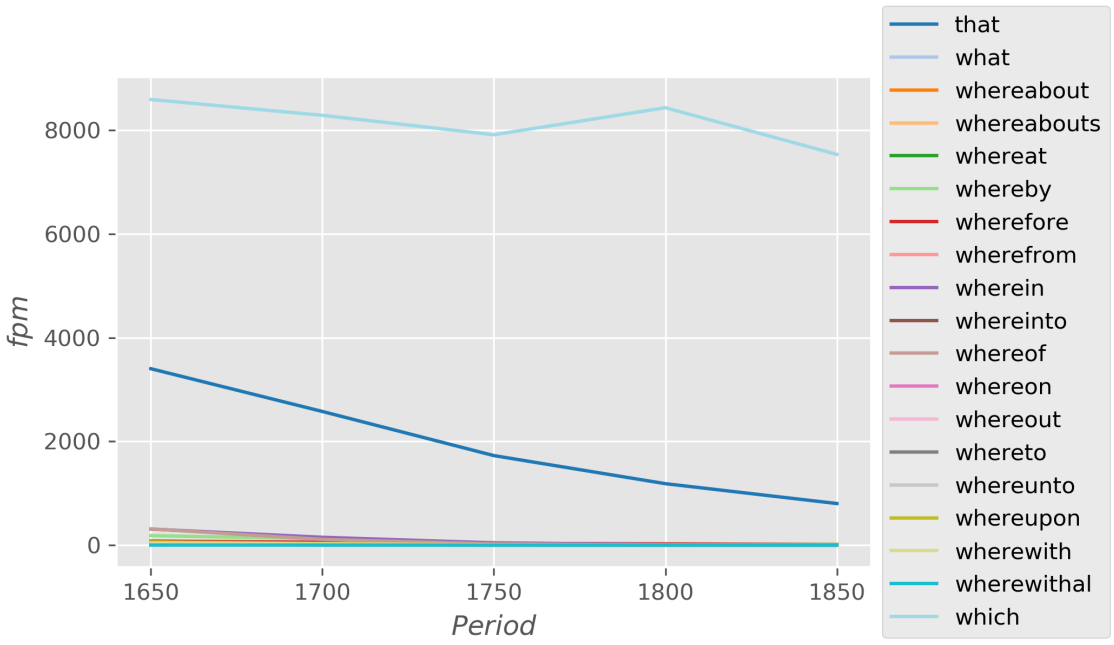


Figure 6.5: Relative frequencies of relativizers per 1 million words (fpm) in scientific English (RSC).

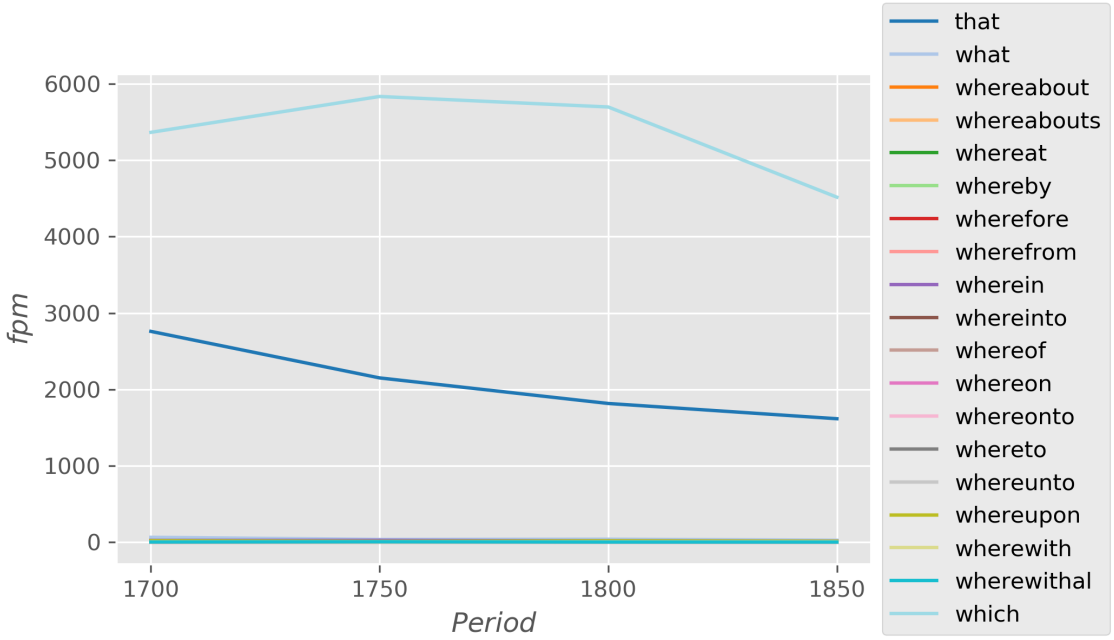


Figure 6.6: Relative frequencies of relativizers per 1 million words (fpm) in general English (CLMET).

Looking at the relative frequencies (frequencies per million tokens; Figure 6.5), we see that *which* decreases only slightly over time, while *that* drops remarkably and the generally very infrequent PAs almost disappear. As discussed in Section 2.3.4.1, the abandonment of PAs does not come as a surprise. The gradual decrease of PAs in the

scientific corpus is in line with the observations by Nevalainen & Raumolin-Brunberg (2012), who claim that the abandonment of synthetic forms such as PAs is part of a "typological drift from synthetic to analytic" (Nevalainen & Raumolin-Brunberg, 2012, p. 203). Moreover, the result that PAs become strongly disfavored is much in line with our assumptions regarding paradigmatic richness. PAs represent a highly variable portion of the relativizer paradigm, and dispensing with them leads to a notable reduction of paradigmatic richness.

For comparison, we look at the percentage distributions of the different relativizers in general English. Figure 6.4 reveals that in general English, the proportions of different relativizers have changed to a much lesser extent. The most noticeable change takes place regarding the choice between *which* and *that*. While *which* becomes stronger proportionally, *that* decreases in proportion towards 1800 and rises slightly afterwards. The distributions of PAs in general English do not seem to change much and their proportion seems to be rather negligible throughout all time periods, suggesting that PAs only played a minor (if any) role in general English, while representing a distinctive feature of scientific English at the beginning of the Late Modern Period. The relative frequencies in general English (Figure 6.6) show that *which* first increases slightly towards the period of 1750 and drops afterward. The trend of *that* shows a different trajectory, descending throughout all observed 50-year periods. The encountered entropy trends clearly reflect the gradual convergence on a preferred option to encode grammatical relations as shown by the distributional trends in Figures 6.3 and 6.4. The reduction in entropy in scientific English over time is owed to an increased probability for *which* to occur as compared to decreasing probabilities of all other available options leading to a smaller choice of options between the different relativizers. The constant values in the general English corpus derive from the comparatively stable proportional distributions of the different relativizers. The results of our entropy calculations show a clear distinction between scientific and general English, pointing to the development of a register-specific preference of *which* in scientific English, while this distinction in general English does not seem to be as strong, maintaining a more equitable choice between *which* and *that*.

6.3.2 German

Regarding the development of the different relativizer types in German, again, we find a relatively stable distribution of relativizers throughout all time periods (Figure 6.7). Similar to English, scientific German shows a preference for two main relativizers: *d.** being the overall most plentiful, followed by *welch.**. Also, *was* takes up a notable proportion. As expected, PAs have a relatively fixed proportion (even increasing slightly) in scientific German until the period of 1800, representing a diverse set of options to introduce RCs in an explicit way. Interestingly, in the last time period (1850–1899), PAs fade out of the picture. This development suggests that a reduction in choice also happens in scientific German, only a century and a half later than in

English. The decrease of PAs seems to be accompanied by an increase of *welch**. The relative frequencies (Figure 6.9) show that the frequency of *d.** peaks in 1750–1800 and decreases afterward, while *welch.** continuously increases in frequency over time.

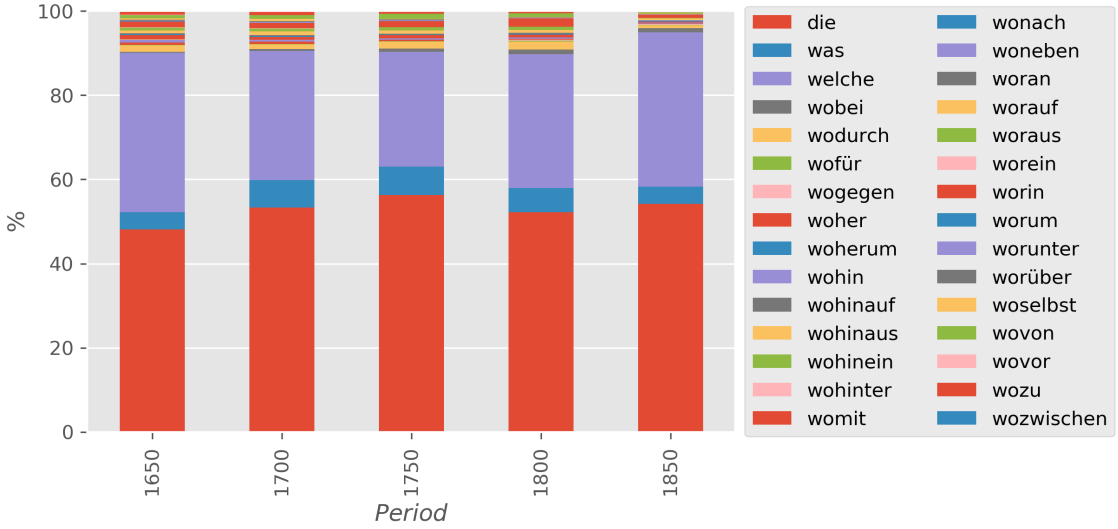


Figure 6.7: Percentage distribution of relativizers in scientific German (DTAW). *die* is the lemmatized form of all instances of *d.** and *welche* represents all instances of *welch.**.

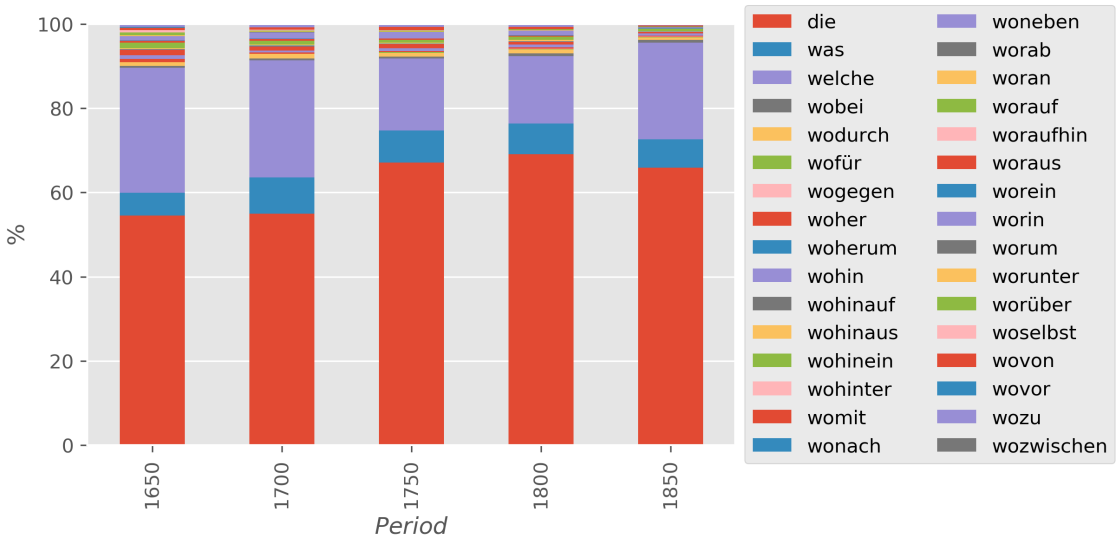


Figure 6.8: Percentage distribution of relativizers in general German (DTAG). *die* is the lemmatized form of all instances of *d.** and *welche* represents all instances of *welch.**.

In the relativizer distributions in general German (Figure 6.8) we see a clear and even increasing preference for *d.**. The proportion of *welch.** instead decreases remarkably until the period of 1800, only to slightly increase again in the period of 1850.

The distribution of *was* is stable over time. The distributional development of the relativizers in general German is similar to that of scientific English in that it shows a similar gradual decrease of PAs. However, general German differs remarkably from scientific German in its gradual decline in the use of *welch.** and its clear preference for *d.**.

The distributions in the German corpora show a different development than those in English. In German, it is the general language that seems to converge on one preferred relativizer option (i.e. *d.**), rather than scientific writing, which throughout the first four time periods seems to uphold a greater paradigmatic richness of relativizers, and only in the last time period (1850–1899) decreases in paradigmatic richness. This reduction is due to the abandonment of many PAs. This reversal trend indicates a time-shifted turn towards lower lexico-grammatical complexity on the level of paradigmatic richness and a turn towards higher processing ease due to lower uncertainty at the choice point of the relativizer.

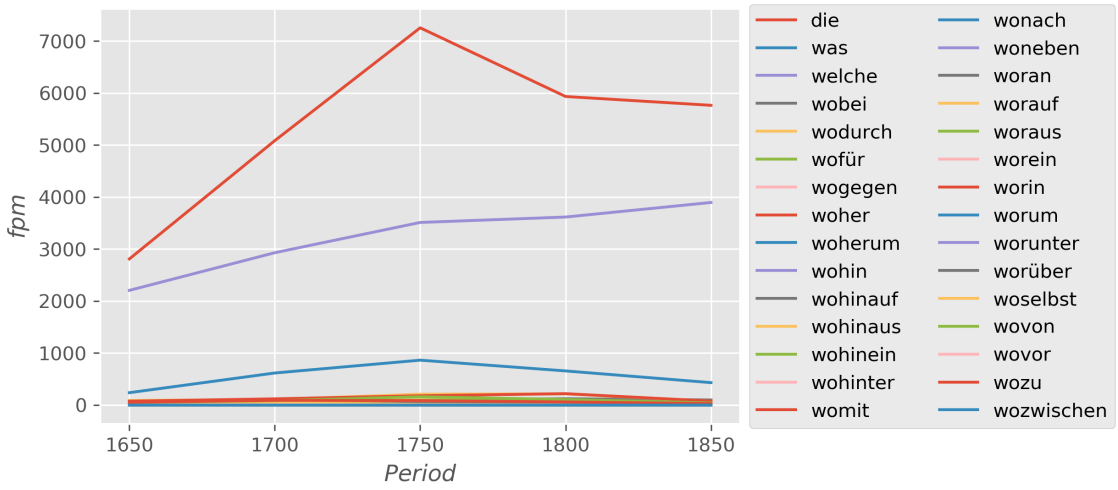


Figure 6.9: Relative frequencies of relativizers per 1 million words (fpm) in scientific German (DTAW). *die* is the lemmatized form of all instances of *d.** and *welche* represents all instances of *welch.**.

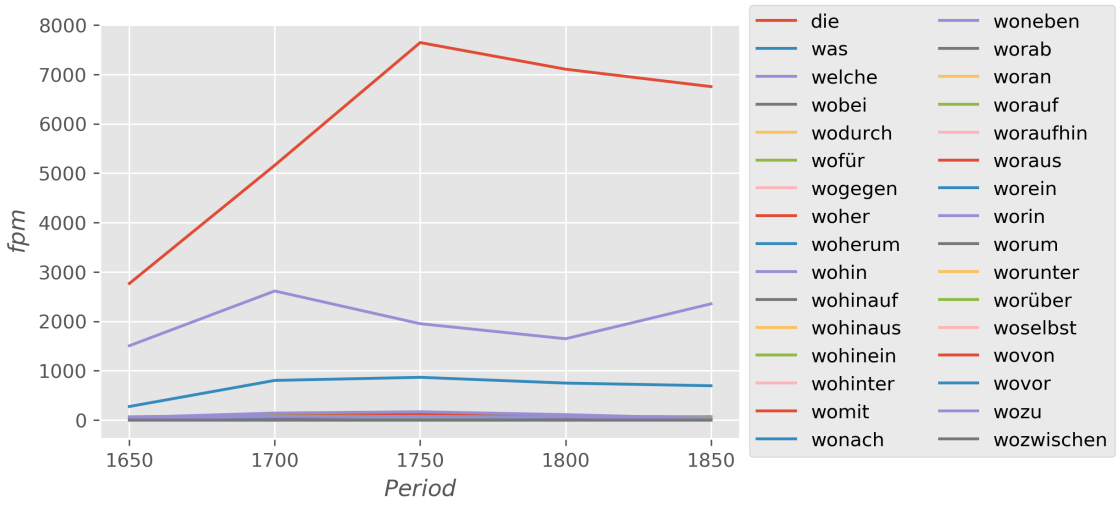


Figure 6.10: Relative frequencies of relativizers per 1 million words (fpm) in general German (DTAG). *die* is the lemmatized form of all instances of *d.** and *welche* represents all instances of *welch.**.

For general German, we found an even greater variety of relativizers (two more relativizer types) than for scientific German, which could be the reason for the slightly higher entropy in the first period. Compared to general English, general German is marked by a steady choice reduction and preference for *d.**, possibly pointing to an increasingly marked stylistic distinction between scientific and general texts. While general German seems to develop a preference for a neutral relativizer (i.e. *d.**), scientific German first shows an inclination towards a rich selection of expressive means, to later on settle on two main options (*d.** and *welch.**). The frequency distributions in scientific (Figure 6.7) and general German (Figure 6.8) show that scientific German increasingly establishes *welch.** as an alternative relativizer to *d.**, while general German increases and consolidates the strong preference for *d.**. We can derive from this that it is actually general German that shows a stronger trend toward choice reduction than scientific German, which tends rather to uphold its diversity of relativizers.

The rise in frequency and variety of relativizers in scientific German until the period of 1800 is consistent with entropy, which reflects increasing complexity regarding relativizer use, and a drop thereof afterward. The stronger tendency of scientific texts towards diversity in relativizer choice during the time between 1650 and 1849 is also in line with Admoni (1990) and Habermann (2011), who report on an initial expansion of grammatical complexity in the German scientific meta-register.

6.4 Summary

Over time, paradigmatic richness in English as measured by entropy seems to be clearly reduced in the scientific meta-register, contributing to the expected trend of complexity reduction on the level of lexico-grammar. The reduction is driven by a strongly conventionalized preference for one relativizer (*which*). At the same time, this trend was not encountered (or much less so) for the general English corpus. We can thus confirm our H1 for scientific English showing a distinctive reduction in paradigmatic richness compared to general English.

The findings for scientific German compared to general German do not confirm our H1, since general German shows an earlier and more straightforward development towards lower paradigmatic richness in terms of entropy and strongly conventionalized preference for one relativizer (*d.**). Scientific German instead preserves a distinctively higher use of *welch.** as a highly frequent alternative to *d.**.

The central role of PAs (being a diverse group of relativizers) in the reduction of paradigmatic richness could be confirmed and seems to be at work in both languages, albeit at different points in time. Contrasting our findings for scientific English and German, we can confirm our H2 regarding language-specific contrasts, since scientific German shows the expected time-shifted trend with an initial increase in paradigmatic richness throughout the first 200 years and a decline afterward.

Chapter 7

Syntagmatic Predictability

As described in Chapter 6, we have discovered that in line with our expectations, English scientific writing has converged on one relativizer (i.e. *which*) being used remarkably more than all other relativizers and much more so than in general English. In scientific German, we found a remarkably higher usage of *welch.** than in general German, where *d.** is strongly preferred. This outcome can guide us in the present analysis looking at the syntagmatic predictability of RCs since we can expect lower surprisal (i.e. higher predictability) of these relativizers as compared to those becoming less preferred in a register. We analyze the surprisal of RCs to occur given their preceding contexts by calculating the 4-gram surprisal (described in Section 5.1.2) on the introductory marker, the relativizer. For this, we first compute the aggregate average surprisal of all relativizers and compare it to the aggregate average surprisal of all words in each corpus per 50-year period (Section 7.1). This comparison is used to address the problem of comparability of surprisal across time due to having different vocabulary and corpora sizes in each 50-year period (also discussed in Section 5.1.2). As more words are introduced to the vocabulary, probabilities are distributed across a larger set. All else being equal, on average, this should lead to a higher surprisal, as individual words become less likely. Using a comparative figure such as the average surprisal of all words in the corpus, we are able to see how the average relativizer surprisal evolves in comparison to the average surprisal of all words in a period. The underlying rationale is that if, due to drastic changes in vocabulary size, surprisal values are subject to major fluctuations, then this should affect the surprisal trend for all words in the same way as the surprisal of a particular group of words (such as relativizers in our case). Hence, we can observe whether relativizers as a paradigm show a distinct development in their predictability in context and whether this development can be regarded as an independent development. In Section 7.2, we inspect the surprisal of different relativizer groups. For English, we define three groups (*which*, *that* and PAs), and for German we group all instances of the lemma *welche* into the group *welch.**, all instances of the lemma *die* into the group *d.** and all pronominal

adverbs into the group PAs. In this analysis, we want to detect individual changes in the predictability of each of the relativizers given their preceding context. Based on our results in Chapter 6, we know that in both scientific German and English, one specific relativizer becomes distinctive over time. We now want to find out whether the paradigmatic shift towards a preferred option is also reflected in increased contextual predictability, i.e. lower surprisal. The overarching assumption here is that we expect the left contexts of RCs should become conventionalized and therefore lead to better predictability of the upcoming RC. This would improve processing, in that a reader of scientific text would have to spend less expectation-based processing load on predicting a syntactic event, leaving more processing resources available for other, e.g. lexico-semantically demanding processing tasks. In Section 7.3, we conduct a qualitative analysis inspecting the most frequent grammatical and lexical contexts (preceding POS and lexical trigrams) of the relativizers *which* and *welch.**.

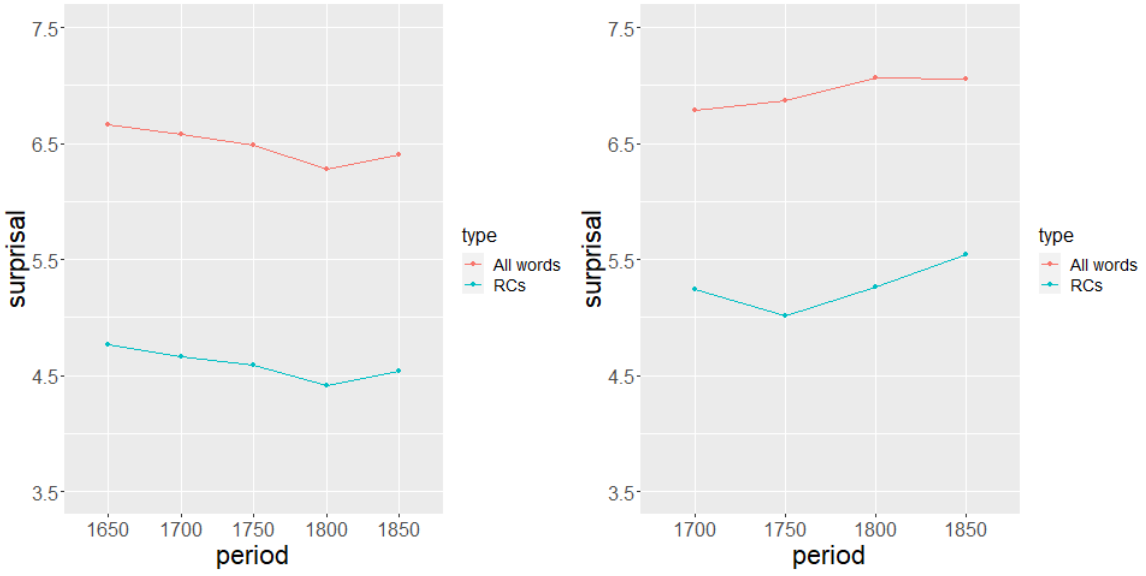
7.1 Syntagmatic predictability of relative clauses

We start by analyzing the development of the overall predictability of RCs given their preceding contexts assuming that RCs overall should become more predictable (i.e. less surprising) over time. To overcome the comparability issue between 50-year periods, we determine diachronic shifts in surprisal by comparing surprisal values of items within a period and measuring the difference between differences, i.e. the difference in surprisal between *ITEM A* and *ITEM B* within one period (e.g. 1650) compared to the difference between *ITEM A* and *ITEM B* in another period. To compare the aggregate surprisal of all relativizers per period, we choose the overall average surprisal (including all words per period) of a 50-year period as a value of comparison to see how the predictability of RCs changes over time.

7.1.1 English

Figure 7.1a shows the average surprisal of all words (red line) vs. the average surprisal of all relativizers (blue line) representing the onset of all RCs in scientific English per 50-year period. Figure 7.1b shows the same for general English. Scientific English shows the same average surprisal trends for all words and for relativizers: a fairly straight downward trend until the period of 1800 and an upward trend afterward. The result is rather unexpected since, due to a constantly growing vocabulary size (compare the strong increase in Types in Table 4.1) in the scientific English corpus, surprisal would be expected to rise (if the number of different words in a corpus grows, the probability of each word shrinks, leading to higher surprisal per word on average). However, this does not seem to be the case. Rather, the conditional probabilities seem to grow and indicate an increasingly conventionalized usage of words, i.e. increasingly similar contexts on which the conditional probability of each word is calculated leads

to higher probabilities and thus lower surprisal. This mechanism seems to be at play for both *all words* and *RCs*.



(a) RSC **(b)** CLMET

Figure 7.1: Average surprisal vs. RC surprisal in (a) scientific (RSC) and (b) general (CLMET) English by 50-year periods.

In general English, the opposite seems to be the case. Figure 7.1b shows an upward trend in surprisal, both for all words and for relativizers, indicating an overall trend toward lower contextual predictability. Part of the rise in average surprisal could derive from an increasing vocabulary size (compare the growing vocabulary size in the CLMET as shown by the number of types in Table 4.6). However, comparing the results to those in scientific English, where the same would apply, we can assume that the growing average surprisal in general English is at least partly due to a diversification of contexts of both RCs and all words in general. Moreover, the growth in vocabulary size in the scientific corpus (RSC) is much stronger than that in the general corpus (CLMET), which would lead to a stronger surprisal increase in scientific English. Instead, the opposite is the case, making an even stronger case for the conventionalization of contexts in scientific English.

To be able to further analyze the specific development of RC predictability, we calculate the differences between the average surprisal values of all words vs. RCs for each 50-year period in both corpora (Table 7.1). The results for scientific English indicate that the differences in surprisal become smaller over time (Table 7.1, column "Difference (avgS)"). The decreasing difference is due to the slightly stronger decrease in average surprisal for all words (Table 7.1, column "All Words (avgS)") than that for RCs (Table 7.1, column "RCs (avgS)"). The average surprisal of all words decreases from 1650 (6.67 bits) to the period 1850 (6.4 bits) by 0.27 bits. The decrease for RC average surprisal from 1650 (4.77 bits) to the period 1850 (4.54 bits) amounts to 0.23

bits. This points to the fact that RCs as a phenomenon do not become exceptionally more predictable over time.

Period	All Words (avgS)	RCs (avgS)	Difference (avgS)
1650	6.67	4.77	1.90
1700	6.58	4.67	1.91
1750	6.49	4.59	1.90
1800	6.28	4.41	1.87
1850	6.40	4.54	1.86

Table 7.1: Average surprisal (avgS) values for all words and RCs and the differences between both values per 50-year period in scientific English (RSC).

In general English (Table 7.2) the differences between average surprisal values for all words and all RCs become smaller over time, too. This is due to the fact that the average surprisal of all words rises by 0.28 bits from 6.77 bits in the period 1700 to 7.05 bits in the period 1850, while the average surprisal for RCs rises slightly more, i.e. from 5.25 bits to 5.55 bits, a difference of 0.3 bits. Thus, in general English, RCs seem to become comparatively more surprising than all words over time.

Period	All Words (avgS)	RCs (avgS)	Difference (avgS)
1700	6.77	5.25	1.52
1750	6.87	5.02	1.85
1800	7.07	5.26	1.81
1850	7.05	5.55	1.50

Table 7.2: Average surprisal (avgS) values for all words and RCs and the differences between both values per 50-year period in general English (CLMET).

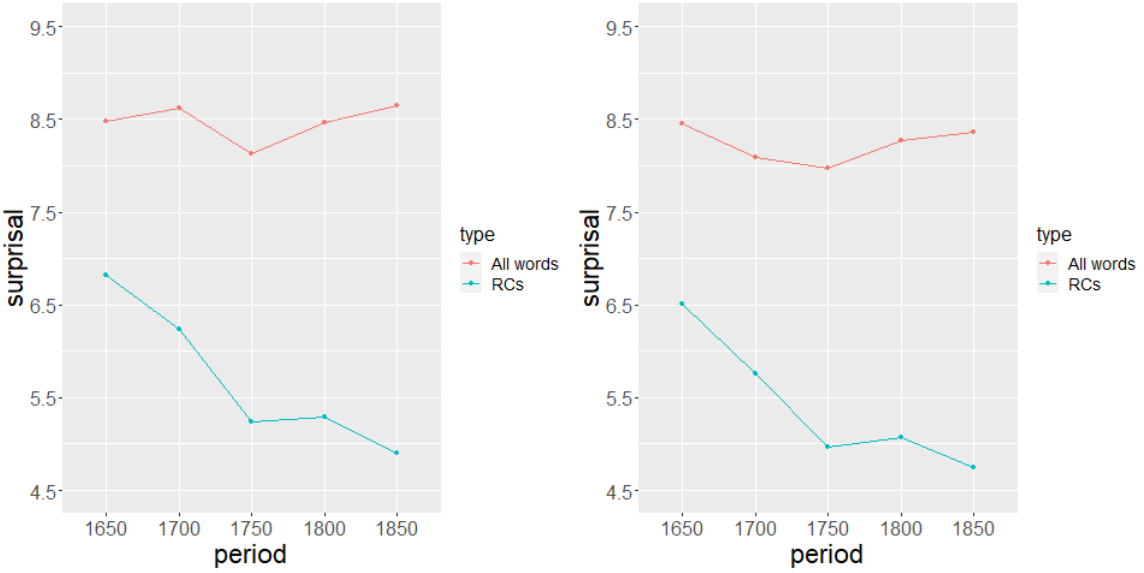
The results show that a growing vocabulary size over time in both English corpora does not seem to have an influence on the trends in surprisal, since in the scientific English corpus the vocabulary growth is much stronger and the surprisal decrease is much lower, while in general English the exact opposite is the case. In terms of register formation, the analysis has therefore confirmed our assumption that RCs in scientific English become less complex over time compared to general English (H1, Section 3.1).

The fact that in scientific English RCs become less surprising also specifically confirms our hypothesis H1.2a (Section 3.1.1.2) that we will find lower surprisal at the onset of RCs. The comparison between all words' and RCs' surprisal in scientific English has shown, however, that the surprisal of all words goes down more strongly than that of RCs. This means that in comparison to the overall surprisal in the

scientific corpus, RCs do not become exceptionally less surprising. The fact that they do become less surprising over time calls for a more fine-grained analysis to find out which syntagmatic contexts become conventionalized and lead to the overall decline of the average surprisal. In general English, we found an opposite trend: The average surprisal of RCs increases more strongly than the average surprisal of all words, pointing to the fact that RCs increasingly occur in more variable contexts.

7.1.2 German

In German, we see an entirely different development than in English. While the average surprisal of all words shows an overall rather stable trend, the surprisal of RCs decreases steeply. This development is remarkably similar in both scientific (Figure 7.2a) and general (Figure 7.2b) German. The contrasting patterns observed between the surprisal values of all words (red line) and RCs (blue line) indicate that the surprisal values are not affected by differences in vocabulary size, i.e. had the vocabulary size had an impact on the surprisal trends, we would have observed parallel trends in the same direction.



(a) DTAW **(b)** DTAG

Figure 7.2: Average surprisal vs. RC surprisal in (a) scientific (DTAW) and (b) general (DTAG) German by 50-year periods.

In addition, there is a remarkable difference in the vocabulary sizes between the German corpora. The scientific German corpus (Table 4.3) shows a significant increase in vocabulary, whereas the general German corpus (Table 4.4) displays a decrease over time. The decrease in vocabulary size observed in general German may seem unexpected, but it can be explained by the gradual standardization of spelling, which leads to a reduction in the number of unique word forms towards standardized

forms. The explanation for the significant growth in vocabulary size in scientific German can be attributed to the expansion of technical terminology in this domain. The surprisal trends of RCs in both scientific German corpora, therefore, show that RCs seem to become highly predictable in their respective contexts. Due to the strong decrease in the average surprisal of RCs over time, the differences between the average surprisal of all words and that of RCs increase remarkably.

Period	All Words (avgS)	RCs (avgS)	Difference (avgS)
1650	8.49	6.83	1.66
1700	8.62	6.24	2.38
1750	8.13	5.24	2.89
1800	8.47	5.30	3.18
1850	8.65	4.90	3.76

Table 7.3: Average surprisal (avgS) values for all words and RCs and the differences between both values per 50-year period in scientific German (DTAW).

Table 7.3 shows that the average surprisal of all words in scientific German (Table 7.3, column "all words avgS") grows slightly (by 0.16 bits, i.e. from 8.49 in 1650 to 8.65 in 1850) whereas, in general German, surprisal of all words decreases slightly (Table 7.4, column "all words avgS") by 0.9 bits from 8.45 bits in 1650 to 8.36 in 1850. At the same time, in scientific German, the average surprisal of RCs declines dramatically, by almost 2 bits, from 6.83 to 4.9. In general German, the decline is not quite as sharp (1.76 bits). The growth in average surprisal of all words and the comparatively stronger decrease in surprisal of RCs in scientific German leads to an overall stronger divergence of the average surprisal trends. In scientific German the divergence ("diff avgS") grows by over 2 bits (from 1.66 in 1650 to 3.76 in 1850). In general German, the divergence grows by 1.67 bits.

Period	All Words (avgS)	RCs (avgS)	Difference (avgS)
1650	8.45	6.51	1.94
1700	8.09	5.76	2.32
1750	7.98	4.97	3.01
1800	8.27	5.07	3.20
1850	8.36	4.75	3.61

Table 7.4: Average surprisal (avgS) values for all words and RCs and the differences between both values per 50-year period in general German (DTAG).

The results indicate that in scientific German, the surprisal of RCs seems to decrease more when observed relative to the overall surprisal in the corpus and when

compared to the trends in general German. The results are plausible since we expect a stronger trend toward higher contextual predictability of RCs in scientific German than in general German due to an increasingly conventionalized usage of RCs over time. Next, we analyze the surprisal of the different relativizers and their particular contributions to the overall downward trend of RC surprisal in the scientific corpora.

7.2 Syntagmatic predictability of specific relativizers

In light of our previous discovery in Section 7.1 that the average surprisal of RCs in the scientific corpora exhibits a declining trend over time, we will delve deeper into this phenomenon and explore the individual contributions of the different relativizers to this trend. Specifically, we will compare the surprisal values of three groups of relativizers: *which*, *that* and PAs in English, and *welch.**, *d.** and PAs in German. To do so, we will inspect the diachronic trends for each relativizer group. As in the previous analysis, we have to keep in mind that the surprisal trends may be biased by the different vocabulary sizes in each 50-year period. Thus, as done in Section 7.1, we also compute the differences between two groups (e.g. *which* vs. *that*) within a particular period (e.g. 1650), and contrast these differences with their counterparts in another period (e.g. 1850).

7.2.1 English

We start by analyzing the distributions of surprisal values per relativizer type in scientific English illustrated by box plots (Figure 7.3)[1]. *PA*s continuously and strongly increase in surprisal. Also, *that* shows an upward trend. In contrast, the surprisal of *which* appears to remain stable over time. The decreasing frequency of *PA*s (discussed in Chapter 6 and illustrated in Figure 6.5) may be a contributing factor to their increasing surprisal. Similarly, the decrease in frequency of *that* may be linked to its increasing surprisal. The stable trend of *which* is possibly due to its stable frequencies over time. Overall, we note a slight increasing trend in surprisal for all relativizers. The trend does not reflect the aggregate trend of RC surprisal we found in Section 7.1.1. It can, however, be explained by the fact that over time, the less surprising relativizer *which* becomes most frequent and thus exerts the strongest influence on the overall trend.

Regarding the bias by the growing vocabulary size in the corpus, the assumption is that the conditional probabilities of the relativizers are calculated on the basis of a bigger vocabulary size and will thus result in lower probabilities overall and

[1]The numbers in the boxes in Figures 7.3, 7.5, 7.7, and 7.9 represent the average (mean) surprisal per group and period. We added the mean values to control whether mean and median values reflect the same trends.

subsequently lead to overall higher surprisal values. Such changes could account for the shared upward trend in surprisal for all relativizer groups. Figure 7.3, however, shows that the trend of *which*, especially, is different from those of *that* and PAs.

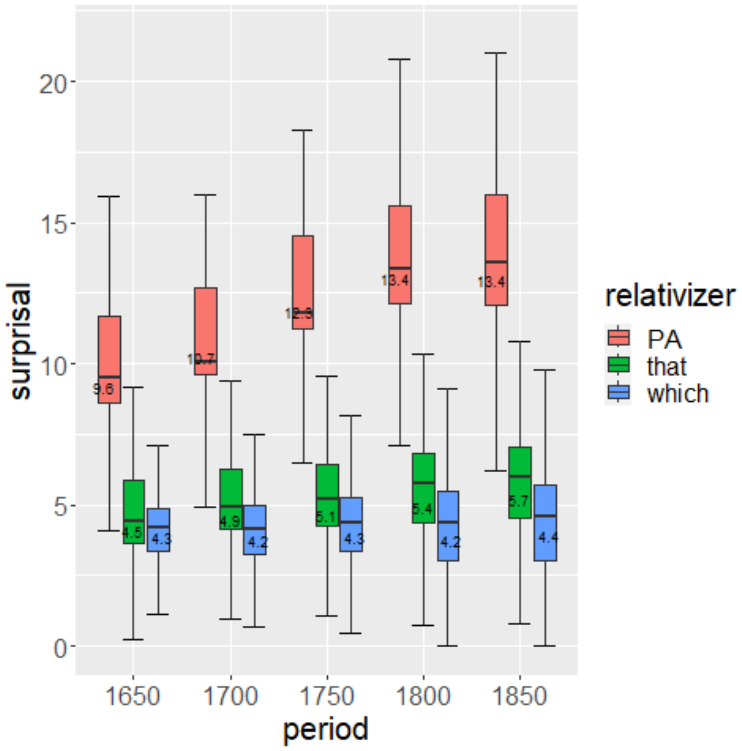


Figure 7.3: Distributions of surprisal values and average surprisal (numbers in boxes) of *which*, *that* and *PAs* per 50-year period in scientific English (RSC).

Comparing the surprisal values in each time period to each other gives us an approximation to interpreting surprisal trends between the three groups over time. To do so, we calculate the average difference between the median surprisal of two relativizers in one period. Figure 7.3 shows that the median surprisal of the three groups seems to steadily grow apart. We thus calculate the average distance between the median surprisal values for each relativizer in one period, e.g.

$$\frac{(\mathrm{S}_{PA} - \mathrm{S}_{which})^{1650} + (\mathrm{S}_{PA} - \mathrm{S}_{that})^{1650} + (\mathrm{S}_{which} - \mathrm{S}_{that})^{1650}}{3} \tag{7.1}$$

A time-series analysis fitting a linear regression model with time as the predictor variable and the mean difference in differences between the median surprisal of the three relativizer groups yields a significant p-value (F-statistic: 47.28 on 1 and 3 DF, p-value: 0.00639) suggesting that the alternative hypothesis that the mean difference between median surprisal increases steadily over time can be confirmed. The linear relationship between time and growth of differences in surprisal is illustrated in Figure 7.4. The red line shows the relationship between time and mean difference, with

time as the predictor variable and mean difference as the response variable. The plot illustrates that the mean difference between the median surprisal of all relativizer groups increases as time increases, indicating a positive linear relationship between the two variables.

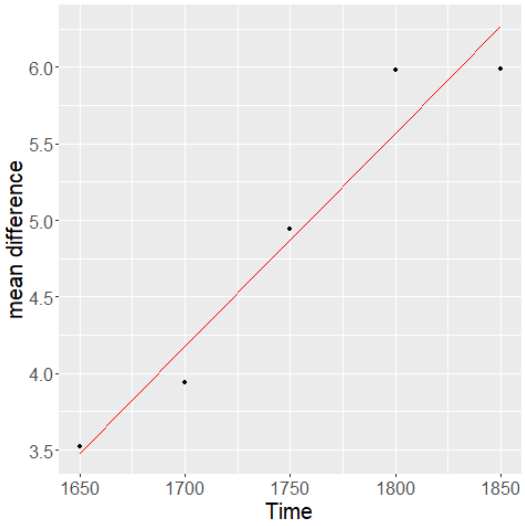


Figure 7.4: Mean differences between the three relativizers groups per 50-year period with a fitted linear regression line in scientific English (RSC).

We can conclude from this that the differences in surprisal between *which*, *that*, and *PAs* increase significantly over time. The differences are due to the fact that the median surprisal of *which* stays fairly stable, but the median surprisal of *that*, as well as for *PAs*, increases in relation to *which*. This increasing divergence underlines the fact that *which* becomes established as the comparatively least surprising relativizer option, while the other options become increasingly unpredictable, hence more surprising. The result is in line with our entropy analysis (Section 6), which has shown that *which* increasingly becomes the preferred option amongst all relativizers. The surprisal analysis has now added the insight that also in context, *which* becomes comparatively more predictable than the other options. In the qualitative analysis (Section 7.3), we will further investigate the grammatical and lexical contexts that *which* settles in.

General English (Figure 7.5) shows a less pronounced development than scientific English. The trend of the three relativizer groups is generally upwards, suggesting that the underlying vocabulary size per period may play a role in the overall surprisal trend. Like in scientific English, *PAs* are the most surprising relativizer type, while *that* and *which* display lower and more similar surprisal distributions. As in scientific English, *which* has a slightly lower surprisal than *that*; however, we do not find that the medians drift apart as much as in scientific English, indicating that both relativizers stay similarly predictable. Furthermore, the distances between the median surprisal values of the three relativizer groups are relatively similar between the different time periods, indicating that the predictability of the different relativizer

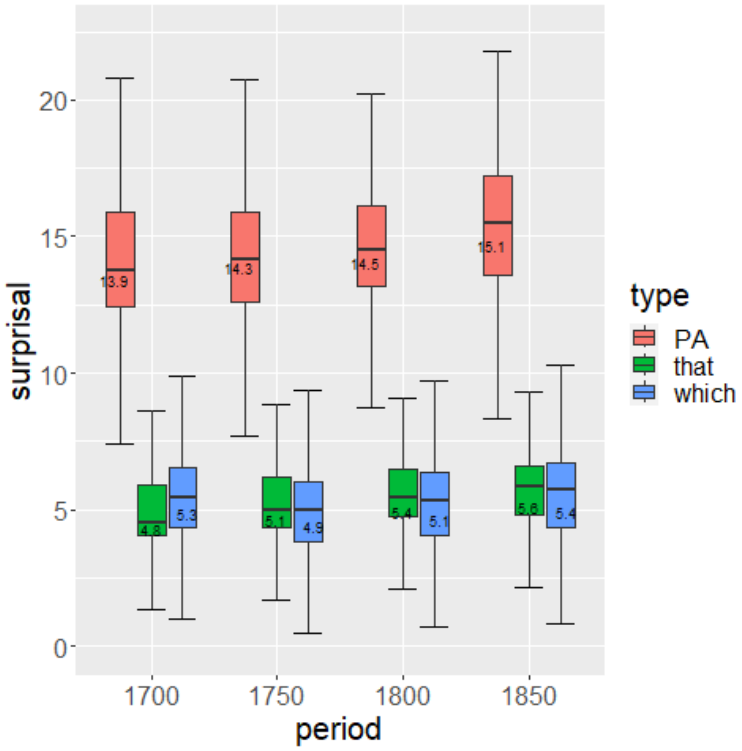


Figure 7.5: Distributions of surprisal values and average surprisal (numbers in boxes) of *which*, *that* and *PAs* per 50-year period in general English (CLMET).

options has not changed dramatically over time, or simply is not interpretable due to variation in corpus size. Looking at the two main relativizers *which* and *that*, we can see that both show similar differences in medians across periods. The time series analysis fitting a linear model with time as the predictor variable and the mean of differences as the response variable also reflects this with a non-significant p-value (F-statistic: 11.09 on 1 and 2 DF, p-value: 0.07956), indicating that differences between the groups do not show a significant trend. Figure 7.6 illustrates the poor fit of the linear model.

The results of this analysis reflect that the differences between the surprisal values of the three relativizers do not change significantly over time since all relativizers collectively become more surprising over time. The results resemble the stable trend in entropy we found in Chapter 6 suggesting that in general English paradigmatic richness and thus predictability at a given choice point, as well as syntagmatic predictability of relativizers, do not seem to change considerably in the Late Modern Period. The finding that *which* seems to gradually become the least surprising relativizer across registers is non-trivial, bearing in mind that *which* is far more common in scientific English than in general English. However, *which* seems to occur in increasingly conventionalized contexts in both registers. In the qualitative analysis (Section 7.3), we will therefore inspect the respective lexical and grammatical con-

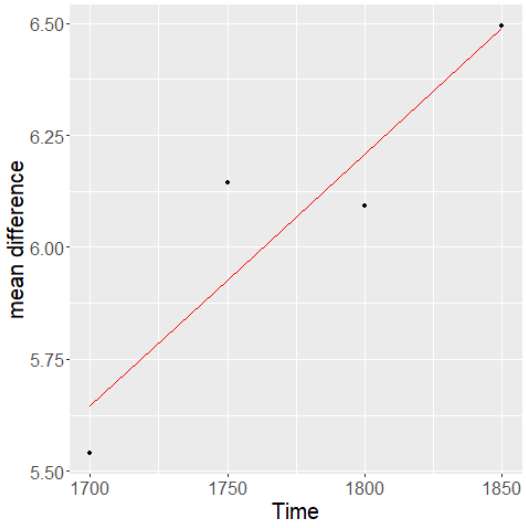


Figure 7.6: Mean differences between the three relativizer groups per 50-year period with a fitted linear regression line in general English (CLMET).

texts of *which* to see the underlying lexico-grammatical structures leading to the comparatively low surprisal in scientific and general English.

7.2.2 German

In scientific German, we find a development opposite to that in scientific English. All three relativizer groups show a downward trend.

This shared downward trend is in line with the general downward surprisal trend of RCs we discovered in Section 7.1. Apart from the fact that surprisal decreases for all individual relativizer groups, the differences between the median surprisal values seem to become smaller over time.

Calculating a time series analysis by fitting a linear regression with time as a predictor of the difference in surprisal between the relativizers over time, we find a significant negative relationship between time and surprisal differences (F-statistic: 10.64 on 1 and 3 DF, p-value: 0.04709) which is reflected in Figure 7.8. Interpreting the observed convergence in median surprisal values on the basis of the individual trends of the relativizer groups (Figure 7.7) it seems that the continuous decrease in surprisal of *welch.** and the stabilizing surprisal values of *d.** and PAs leads to the decreasing difference between the three relativizer groups. The average surprisal values for each group per 50-year period (displayed as numbers in the boxes in Figure 7.7) show that the average surprisal of *welch.** in 1850 (4.5 bits) is even below the average surprisal of *d.** (4.7 bits) although the median of *welch.** is higher than that of *d.**. The lower average surprisal reflects the leftward skewness of the distribution showing that *welch.** increasingly takes very low surprisal values. These results indicate that on the one hand, all three relativizer groups seem to become less surprising in their preceding contexts. It is, however, especially the straight downward trend

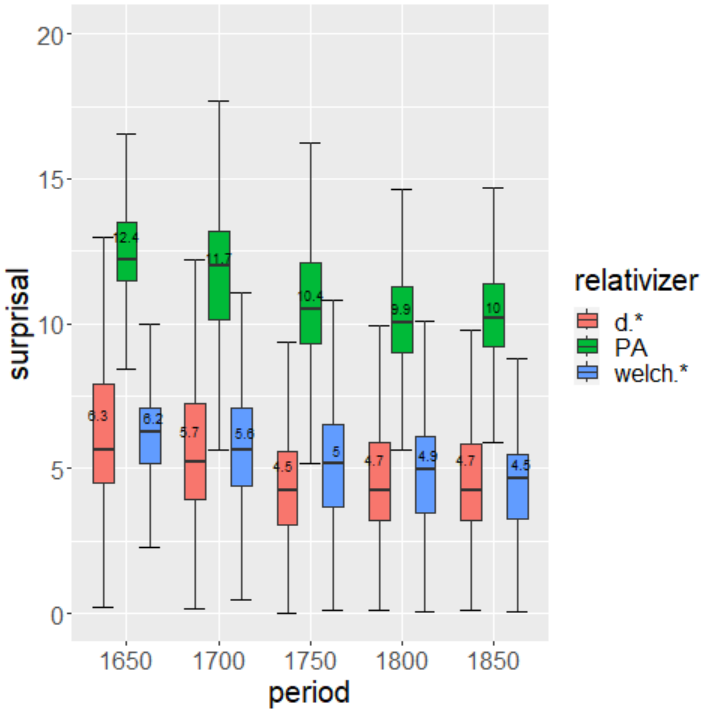


Figure 7.7: Distributions of surprisal values and average surprisal (numbers in boxes) of *welch.**, *d.** and *PAs* per 50-year period in scientific German (DTAW).

in surprisal of *welch.** that suggests that the relativizer settles in increasingly conventionalized contexts. To find out which lexical and grammatical contexts *welch.** settles in specifically, we will inspect them in the qualitative analysis (Section 7.3).

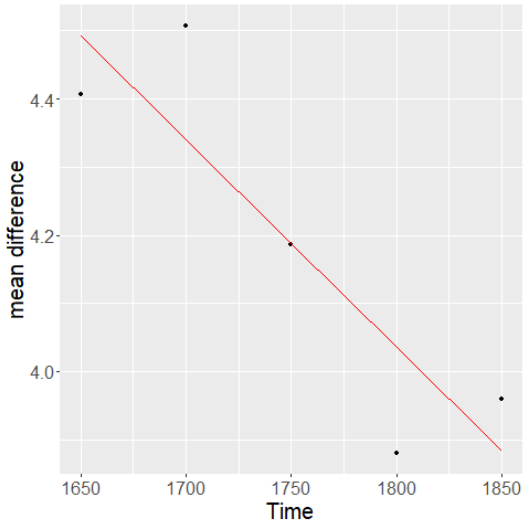


Figure 7.8: Mean differences between the three relativizer groups per 50-year period with a fitted linear regression line in scientific German (DTAW).

Inspecting general German for comparison, we see that the three relativizer groups show a less straightforward trend in terms of surprisal. Despite the finding that RCs overall become less surprising over time as opposed to a fairly stable average surprisal in the corpus (Section 7.1, Figure 7.1b), the development within the groups is not uniform. For instance, the surprisal of *welch.** decreases steadily, as in scientific German, while both *d.** and *PAs* have their lowest median surprisal in 1750 and a mild upward trend after that. The differences between the median surprisal values therefore also do not show the converging trend that we found for scientific German.

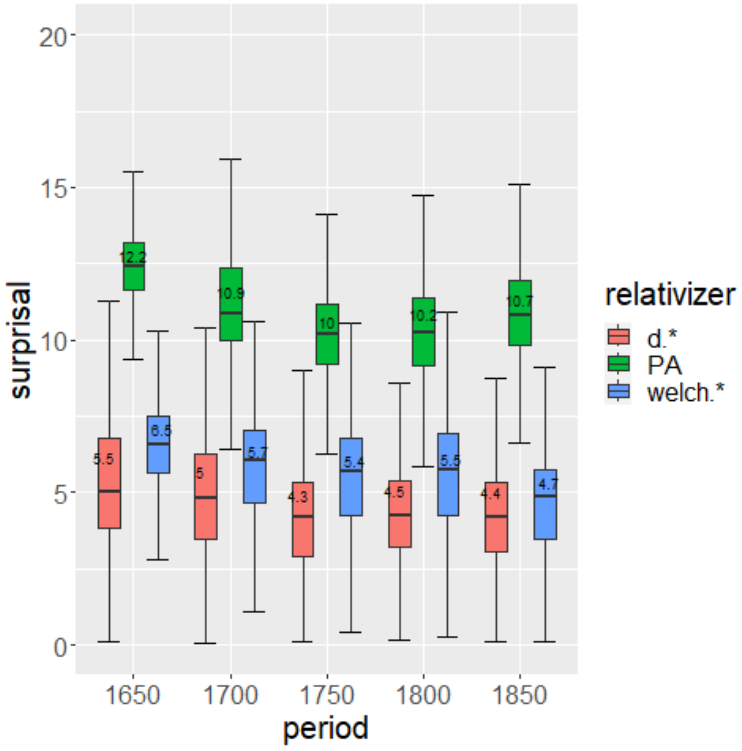


Figure 7.9: Distributions of surprisal values and average surprisal (numbers in boxes) of *welch.**, *d.** and *PAs* per 50-year period in general German (DTAG).

The time series analysis based on a linear model taking time as a predictor variable of the mean difference between the three relativizer groups yields a non-significant negative relationship (F-statistic: 0.5706 on 1 and 3 DF, p-value: 0.5049). The poor fit of the model is reflected in Figure 7.10. We can derive from this analysis that in the 18th c. and the first half of the 19th c., the three groups became more similarly surprising in their contexts than they were in the 17th c. and at the end of the 19th c. At the end of the 19th c., the two main relativizers *d.** and *welch.** become especially similar in surprisal due to a notable reduction in the surprisal of *welch.**. Although surprisal values are not exactly comparable between periods, we can note a continuous downward trend in surprisal for *welch.**, while *d.** first goes down until the end of the 18th c., and then goes up again in the 19th c. Unlike in scientific German, *welch.** has

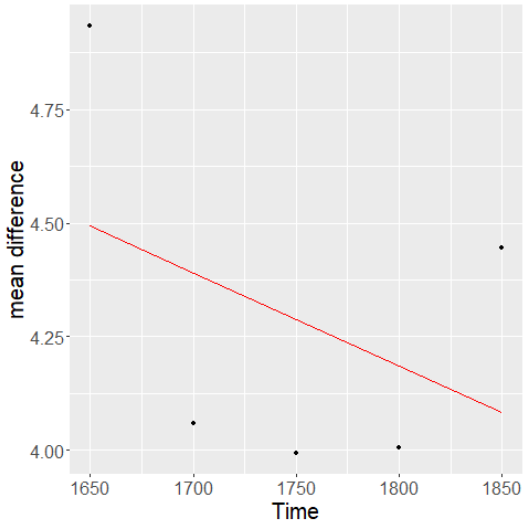


Figure 7.10: Mean differences between the three relativizers groups per 50-year period with a fitted linear regression line in general German (DTAG).

a constantly higher mean and median surprisal than *d.**, indicating that *d.** keeps its position as the most predictable relativizer in general German. The non-linear trend in mean differences in surprisal also suggests that in general German, the relativizer groups do not change in surprisal as dramatically as in scientific German, where *welch.** steadily develops towards higher predictability.

In summary, the results for English and German confirm our hypothesis that scientific writing develops toward increasing syntagmatic predictability of RCs in general (H1.2a, Section 3.1.1.2) and of certain relativizers in each register and language (H1.2b, Section 3.1.1.2). While in scientific English the overall syntagmatic predictability of RCs increases, the specific relativizer responsible for this increased predictability seems to be the most frequent of the relativizers, namely *which*. In scientific German, all relativizers become more predictable in their contexts and thus collectively seem to contribute to the overall higher predictability of RCs over time. However, also in scientific German, one specific relativizer (i.e. *welch.**) shows an exceptionally strong trend toward lower surprisal. We need to keep in mind that there might be different mechanisms at work motivating the increased predictability in each language. For this reason, we conduct a qualitative analysis by inspecting the respective syntagmatic contexts of the two relativizers that have proven to be the most predictable ones in context.

7.3 Syntagmatic contexts of relativizers

In the ensuing analysis, we perform a qualitative examination that focuses on the relativizers *which* in English and *welch.** in German. This selection is made based on their pronounced trend towards greater syntagmatic predictability, which is particu-

larly prominent in scientific writing. In order to investigate the contexts in which the relativizers *which* in English and *welch.** in German occur and whether these contexts change over time, we analyze both grammatical and lexical contexts preceding the target words. Part-of-speech (POS) trigrams are used to examine the grammatical contexts, while the preceding lexical trigrams are analyzed to study the lexical syntagmatic context. Our analysis begins with an inspection of the grammatical contexts.

7.3.1 Grammatical contexts

To detect the most impactful grammatical contexts of *which* and *welch.**, we examine the three most frequent POS trigrams preceding the relativizers in each time period (e.g. determiner noun adjective + *which*). In each 50-year period, different trigrams may belong to the top three, and we consider the trajectory of each encountered trigram across the whole 250 years. Some of the most common trigrams in a period may (or may not) overlap with those from another period; thus the total number of trigrams displayed in the Figures may vary between the corpora. If the total number of trigrams displayed is low, it suggests a lower level of variation between periods (since it is always the same trigrams that are among the top three), while a higher number indicates a greater level of variation. Our assumption is that grammatical patterns in scientific language become more standardized, i.e. we encounter convergence in the use of particular patterns.

7.3.1.1 English

The trigram contexts in scientific English (Figure 7.11) show an astonishingly clear trend: all but one of the most frequent pos trigrams identified as frequent left contexts preceding *which* decrease in relative frequency[2]. The only POS trigram that almost linearly increases across periods is the pattern [DT NN IN] (determiner noun preposition; compare Example (2)). At the same time, all other trigrams decrease in frequency, and all of them include a comma. For instance, the trigram [NN , IN] may be a variant of punctuation of the preferred trigram [DT NN IN] in the 19th c. Notably, the trigram [DT NN ,] decreases drastically. This may be due to a change in marking defining and non-defining RCs. As we can see in Example (1), in 1665 a non-defining RC was still separated by a comma from the matrix clause, which in present-day English would be a violation of the rules of punctuation.

(1) *But he proceeds to speak of the Inclination, which the Mandril must have upon the Plain of the Ring.* (Monsieur Auzout, 1665)

[2]Most frequent trigrams in **1650** and **1700**: [DT NN ,], [JJ NN ,], [JJ , IN], **1750**: [DT NN ,], [JJ NN ,] and [DT NN IN], **1800**: [DT NN IN], [NN , IN], [JJ NN ,], **1850**: [DT NN IN], [JJ NN ,], [NN , IN].

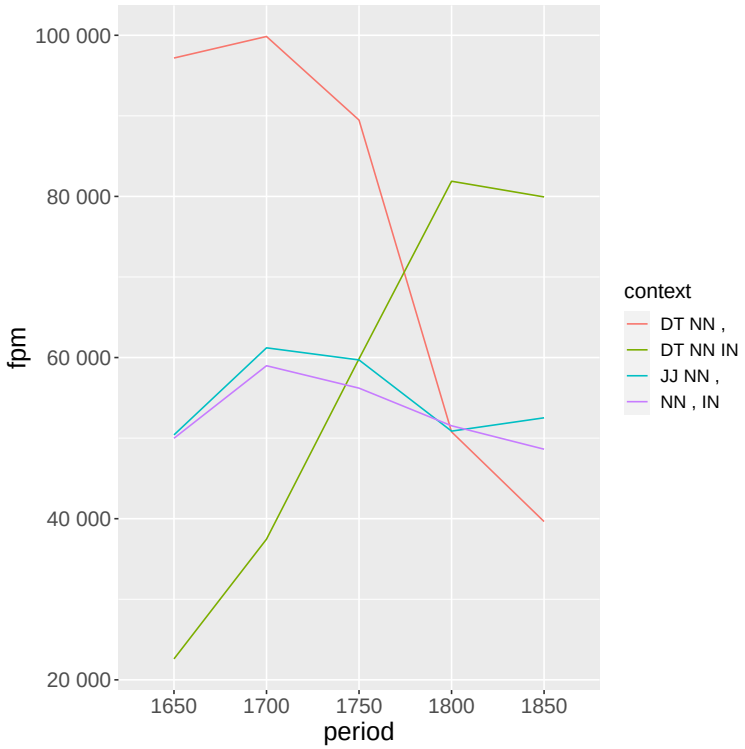


Figure 7.11: Three most frequent part-of-speech trigrams preceding *which* per 50-year period in scientific English (RSC). Trigrams that are among the top three in another period are also included to show the trajectory over the periods; thus the graph shows more than 3 lines.

In general, we observe a decrease in the frequency of previously common contexts for *which* in scientific writing, while one previously less favored pattern becomes the predominant context for RCs: *noun phrase + preposition + which*, as exemplified in Example (2).

(2) [...] *the R. Society,* ***the design of which*** *I admire as the Noblest, that ever was undertaken by men.* (Philosophical Transactions, 1665–1678)

This finding corroborates our assumption that in scientific writing contexts of RCs become increasingly conventionalized in line with Degaetano-Ortlieb & Teich (2019) and Teich et al. (2021). Compared to general English (Figure 7.12), the variation of most frequent POS contexts preceding *which* (i.e. the number of POS trigrams displayed in the figure) is lower in scientific English. For scientific English, the three most frequent trigrams of each period are mostly identical, with the exception that [DT NN IN] takes over the first rank in 1800 and [DT NN ,] drops out of the top three in 1800. In general English, instead, the top three trigrams change entirely over time, i.e. [DT NN,], [NN , IN] and [JJ NN ,] in the first two periods (1700 and 1750) and [DT NN IN], [IN DT NN], and [DT JJ NN] in the last period. Furthermore, among this more diverse set of most frequent contexts in general English, three POS

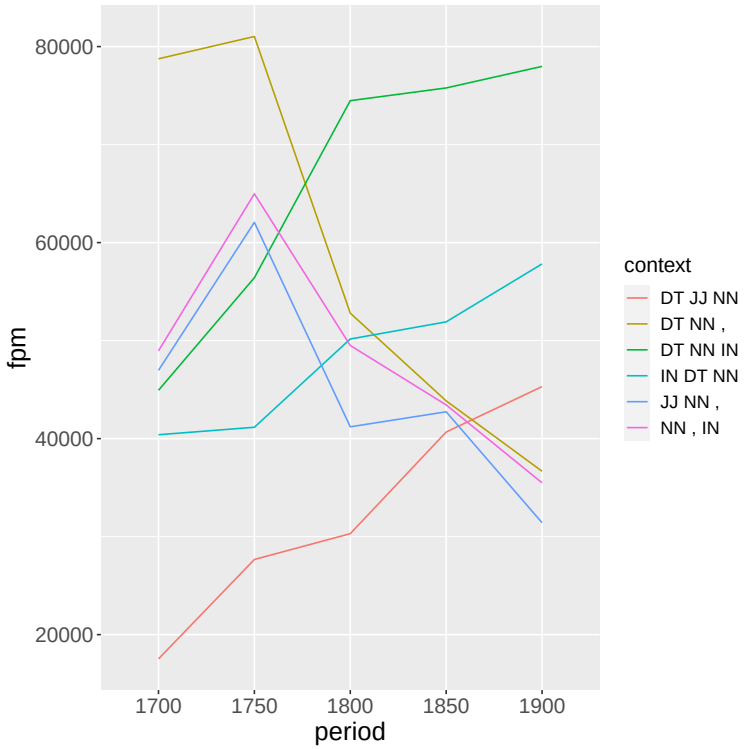


Figure 7.12: Three most frequent part-of-speech trigrams preceding *which* per 50-year period in general English (CLMET). Trigrams that are among the top three in another period are also included to show the trajectory over the periods; thus the graph shows more than 3 lines.

patterns increase in frequency over time, signaling that the grammatical contexts of *which* in general English are more diverse than in scientific English, with only one pattern increasing towards the last period. As in scientific English, the most frequent context is [DT NN IN] (determiner noun preposition), but the patterns [IN DT NN] (preposition determiner noun) and [DT JJ NN] (determiner adjective noun) also become more frequent over time.

In summary, the encountered POS patterns represent typical constructions that introduce RCs in the two corpora. The analysis has revealed the expected increasingly conventionalized usage of RCs in grammatically heavily restricted contexts in scientific writing and more diverse usage of RCs in general English. Since the analysis of grammatical contexts does not directly reflect contexts relevant for lexical surprisal, we will analyze the lexical trigram contexts introducing *which* in Section 7.3.2.

7.3.1.2 German

In scientific German (Figure 7.13), the most frequent POS contexts preceding *welch.** are less diverse than in English, showing three clearly preferred contexts: [ADJA NN PT] (adjective noun comma) representing long premodified nominal phrases (see Example (3-a)), [ART NN PT] (article noun comma) representing shorter nominal

phrases, and [NN PT APPR] (noun comma preposition).

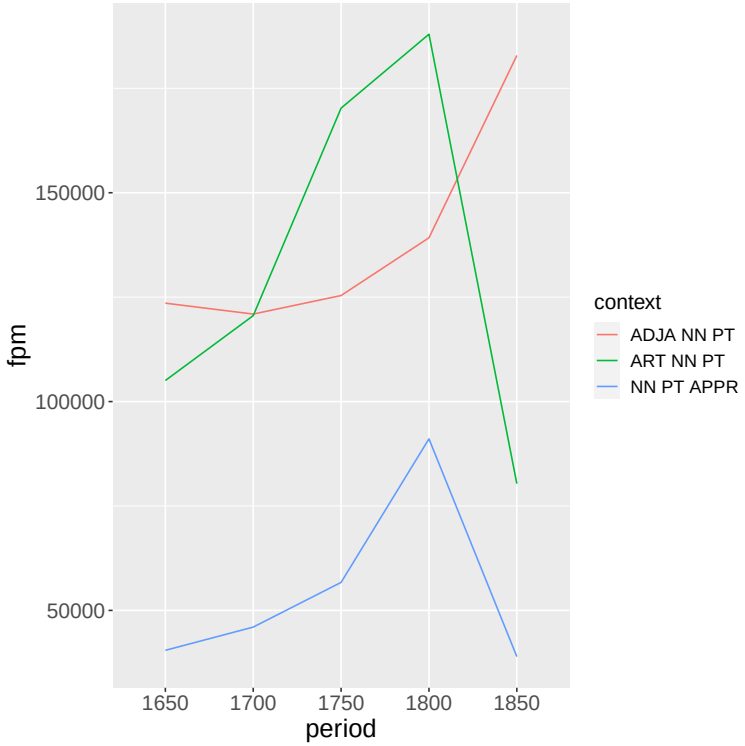


Figure 7.13: Three most frequent part-of-speech trigrams preceding *welch.** per 50-year period in scientific German (DTAW). The three most frequent trigrams are identical in all 50-year periods; thus the graph shows three lines.

(3) a. *[...] so gehet alles Saltz mit einem Theil deß zugesetzten ins Wasser/ vnd gibt eine* ***Graßgrüne solution, welche*** *man filtriren, vnd ein Theil deß Wassers wieder abstrahiren soll [...].* (Johannes Glauber, Opera Chymica, 1658)

b. *Beeren aller Art liebt dieses Huhn ganz ungemein, und ihnen zu Gefallen besteigt es die Wipfel* ***der Gebüsche, welche*** *sie hervorbringen; aber auch Baumfrüchte, z.B. Aepfel, behagen ihm sehr.* (Alfred E. Brehm, Illustrirtes Thierleben, 1867)

In fact, these three contexts are the most frequent contexts in both scientific and general German (Figure 7.14). The fact that these three patterns are continuously among the three most frequent POS contexts in both corpora alike points to a relatively rigid use of RCs in German compared to English. Only the trajectories of their frequency developments are slightly different, especially in scientific German after 1850. While before that all three trigrams increase in frequency in both corpora, after 1850 in scientific German, the two patterns [ART NN PT] and [NN PT APPR] suddenly decrease steeply, while the trigram representing a long noun phrase [ADJA

NN PT] continues to increase in frequency. In general German, all three patterns continue to increase.

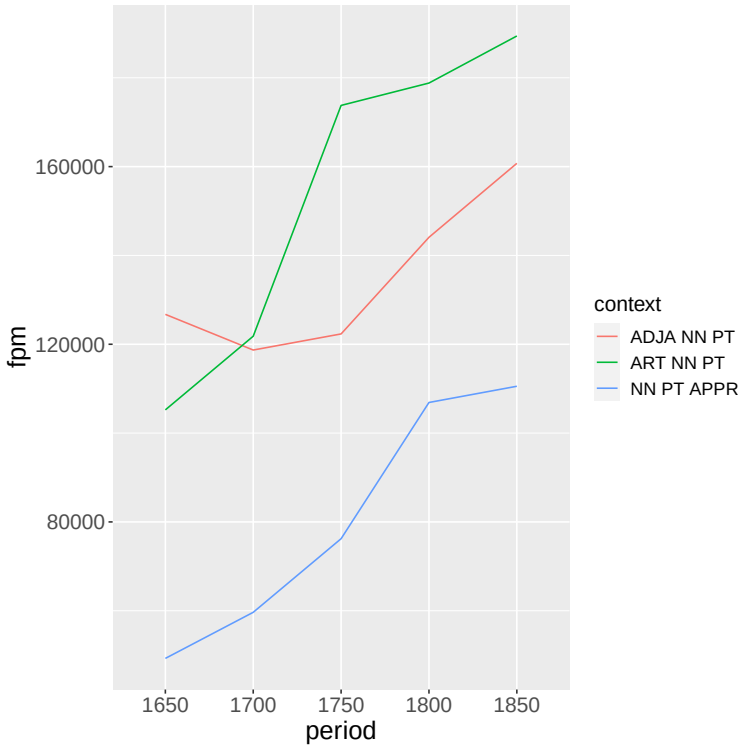


Figure 7.14: Three most frequent part-of-speech trigrams preceding *welch.** per 50-year period in general German (DTAG). The three most frequent trigrams are identical in all 50-year periods; thus the graph shows three lines.

The observed development reflects that in the late 18th c., grammatical contexts of RCs in scientific German become highly restricted to one specific POS trigram, while before that, RC contexts seem to be more diverse. This convergence on one preferred context reduces the grammatical variability of syntagmatic contexts of RCs in scientific German. In contrast to scientific English, scientific German exhibits an initial increase in various grammatical contexts and a time-shifted decrease in grammatical contexts of RCs. This finding confirms our hypothesis that scientific German reduces grammatical complexity later than English after initially showing an increase in grammatical complexity (H2, Section 3.2). The "surviving" grammatical contexts of RCs represent informationally extremely loaded constructions with richly premodified head nouns further postmodified by an RC (see Example (4)).

(4) ***Der rein quantitativen Wirkungsfähigkeit, die*** *wir als physische Energiegröße bezeichnen, lässt sich daher die qualitative Wirkungsfähigkeit in Bezug auf die Erzeugung von Werthgraden als psychische Energiegröße gegenüberstellen .* (Wilhelm Wundt, Grundriss der Psychologie, 1896)

To identify the particular lexical contexts that contribute to the decreasing surprisal of the relativizers *which* and *welch.**, a qualitative analysis will be conducted in the next section.

7.3.2 Lexical contexts

To identify the specific lexical trigrams contributing to the decreasing surprisal of *which* and *welch.**, we conduct a qualitative analysis and examine the three most frequently occurring lexical trigrams that precede the relativizers in the English and German corpora. Since we found that *which* and *welch.** are the most predictable relativizers in scientific language, we assume that they occur in increasingly similar contexts compared to general language. To analyze not only the exact trigram but also the semantic function of the left context of *which* and *welch.**, we summarize the top three preceding lexical trigrams in functional groups.

7.3.2.1 English

Table 7.5 shows the top three lexical trigrams per period in scientific English and the specific functional groups they belong to. The most frequent ones per corpus are highlighted in boldface. Analyzing the trigrams, we can observe that all of the trigrams include a preposition directly preceding the relativizer *which*. From 1750 onward, one specific trigram, *the manner* ***in***, becomes by far the most frequent context preceding *which*. Interestingly, the lexical trigram maps onto the most frequent POS trigram ([DT NN IN]) in scientific English, as we discovered in Section 7.3.1.1. In lexico-grammatical terms, the trigram introduces an RC with an adverbial gap expressing manner (Biber et al., 1999). Syntactically, the construction represents an adjunct with the function of an adverbial. In 1800, *the mode in* adds to the top three trigrams fulfilling the same lexico-grammatical and syntactic characteristics. The other major group of trigrams ends in the preposition *of*. Semantically, most of the trigrams with *of* express quantification (e.g. *, some of / , one of* ; Example (5)).

(5) *The upper part of this animal is covered over with circular cells,* ***one of which*** *is represented at Fig. 4. [...].* (John Ellis, An Account of the Sea Pen, 1763)

However, the semantic notion of quantification only seems to come in at the beginning of the 18th c. Before that, the semantics of the trigrams ending in *of* are less transparent. Interestingly, the type of preposition in the trigrams seems to impact the syntactic characteristics of the trigram: the manner expressing trigrams ending in *in* together with the relativizer *which* form adverbials, while the patterns with *of* syntactically function as genitives. Note that in 1650 the second most frequent trigram is *the doing of (which)*. Also in this case, *which* occupies the syntactic function of a genitive. A look into the corpus reveals that the trigram occurs in highly

conventionalized contexts, and *which* functions as a resumptive pronoun rather than as a relativizer (Examples (6)).

(6) a. *Having dispatched this first Part, he proceeds to the other Part of this Treatise, and therein delivers the History of the Gullet, stomach, and Guts: In* ***the doing of which,*** *he discusseth many considerable Questions; E.g. which Animals have gullets, and which not?*

b. *In the former he endeavours to explain, How the Brain is formed, and what kind of substance it is: in* ***the doing of which*** *he observes [...].*

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

Table 7.5: Lexical trigram contexts of *which* in scientific English (RSC) per 50-year period.

Tracing the trigram through time in the corpus, we find that after 1650 the trigram only occurs three times in 1700 and disappears completely afterward. The texts as well as the topics of the texts that the trigram occurs in are diverse (mechanics, physiology, etc.), which dispels the suspicion that the trigram is an idiosyncratic expression of one specific author. The most obvious corresponding expression to compensate for the loss of the trigram is the substitution pattern *in doing so*. The corpus data reveal that this pattern indeed is a relatively new one, starting to be used only from 1800 onward. Overall, the most frequent trigrams in 1650 are semantically hard to interpret. For instance, *[, out of]* could either be part of a partitive construction as in *"we had ten apples, out of which five were rotten"*, or a prepositional complement functioning as an adverbial of place, as in *"the jar, out of which we took the jam"*. In 1700, patterns of quantification start to arise and conventionalize as the second most

frequent patterns introducing RCs. In 1750, the manner-expressing patterns appear and become the most frequent ones for the next 150 years. Relating our findings from the actual lexical trigrams to our surprisal analysis in Section 7.2.1, the increasing frequencies of the top three patterns in 1750 and 1800 as conventionalized contexts in scientific English could explain the increasingly low whiskers (and first quartiles) of the surprisal values of *which* in Figure 7.3.

In general English, we find that from the earliest period on, manner-expressing trigrams are by far the most frequent trigrams (Table 7.6). We encounter the trigram *the manner in* in the second position as early as 1700. The pattern becomes the most frequent trigram in 1750 and is also most frequent in 1800. Other patterns expressing manner such as *the way in* join the top three in 1800. The latter becomes the preferred pattern introducing RCs with *which* in 1850. A look into the period of 1900 (excluded from analysis in this thesis) reveals that *the manner in* disappears from the top three trigrams and is replaced by *the way in* and *in a way*. The pattern seems to be versatile in the register it can be used in. *The way in* occurs similarly frequently in more formal registers such as treatises (Example (7)) and in less formal registers such as narrative fiction (Example (8)).

(7) [...] *the mathematician, linguist, naturalist, or philosopher, explains* ***the way in which*** *his learning beneficially influences action* [...]. (Herbert Spencer, *Essays on Education*, 1861)

(8) *That was* ***the way in which*** *Miss Amelia reasoned.* (William M. Thackeray, *Vanity Fair*, 1843)

A third trigram becoming highly frequent in the general English corpus is *the sense in*. A closer look at the specific register in which this pattern tends to occur reveals that it is used predominantly in treatises. We may assume that treatises as a register have become similarly conventionalized as scientific texts. Since *the sense in* is a highly explicit way of rendering the meaning of a word or expression, its frequent use in treatises seems obvious.

Comparing the lexical trigrams preceding *which* in scientific and general English, we see that the trigrams most frequently used in English belong to the functional categories of manner and quantification. In the two registers, we see two clear tendencies. In general English, the majority of the top three trigrams (9/15) describe manner expressions, while in scientific English, the trigrams appearing most often are expressions of quantification (8/15). Although quantification is overall more frequent in scientific English, the trigram *the manner in* is the top trigram between 1750 and 1850. In general English, the preferred lexical rendering is *the manner in* until 1800 and then shifts to *the way in*.

The encountered trigrams represent highly conventionalized expressions. Moreover, they are functionally special cases of RCs, since they assume the role of partitives or adverbial RCs rather than defining a semantically rich head noun. Syntac-

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

Table 7.6: Lexical trigram contexts of *which* in general English (CLMET) per 50-year period.

tically, the trigrams introduce genitive RCs or adverbials. Connecting these findings back to our findings for the most common POS trigrams, we see that the manner-expressing lexical trigrams coincide with the most frequent grammatical pattern [DT NN IN] in scientific English. However, the quantification-expressing trigrams, which are the second most frequent trigrams per period, are not among the top three POS trigrams. Note that the trigrams *the manner / mode in* only occupy a small fraction (1395/28739) of the whole set of lexical options of the POS pattern in scientific English, but they constitute the most frequent lexical instances in the set. Appearing in such an accumulated way, the trigram makes the relativizer *which* extremely predictable and accounts for the low surprisal values encountered for *which* in the scientific English corpus.

The fact that the lexical trigram *the manner in* only occupies a small part of the full range covered by the POS trigram [DT NN IN] raises another thought. In Chapter 6, we saw an extreme loss of PAs (Example (9-a)) functioning as relativizers and in the present chapter, Section 7.2.1, we found a strong increase in surprisal of PAs. Many lexical instances matching the POS trigram [DT NN IN] can function as "analytic" forms of the synthetic PAs as shown in Example (9-b).

(9) a. [...] *after the evaporation of a part of* ***the water wherein*** *this salt hath been dissolved* (Philosophical Transactions, 1665–1678)

b. *At the end of eight days,* ***the water in which*** *the albumen had been*

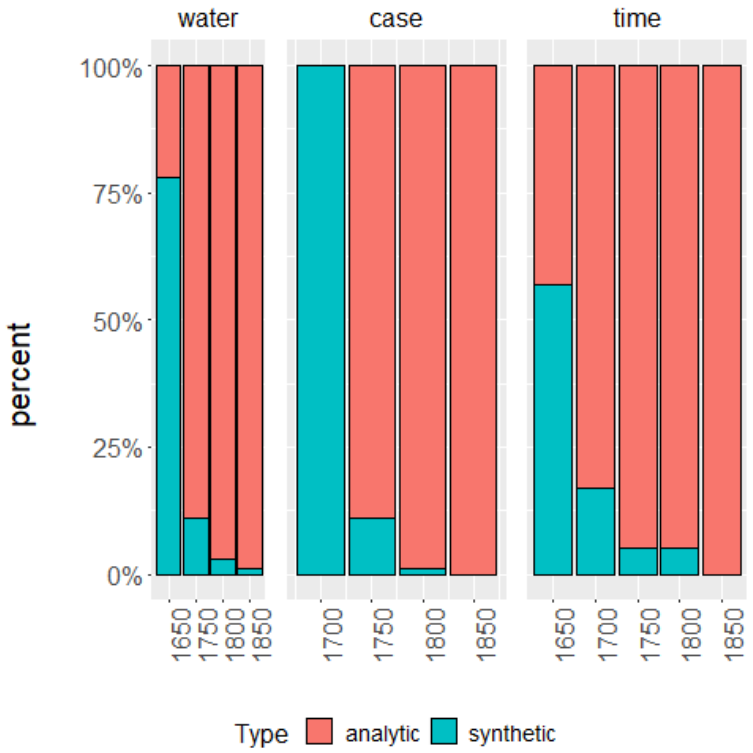


Figure 7.15: Percentage distributions of analytic and synthetic relativization of the target head nouns *water*, *case* and *time* in scientific English (RSC) per 50-year period.

> *digested was examined [...].* (Charles Hatchett, Chemical Experiments on Zoophytes, 1800)

To find out whether there may have been a replacement of the synthetic PAs by analytic forms, we look at the example case of *wherein* and the three most frequent lexical items preceding it. The most frequent items are *water*, *case*, and *time*. For each item, we extract the normalized frequencies of the two corresponding patterns occurring with an analytic relativization strategy, i.e. *water/case/time in which*, and with a synthetic PA, i.e. *water/case/time wherein*. The resulting percentage distributions are displayed in Figure 7.15.

We can see that for all synthetic variants, the frequencies start out higher than those of the analytic variants. In the case of *case in which/case wherein*, we see that the first occurrences are exclusively synthetic in 1700. We also see that the shift from synthetic variants toward analytic variants in 1700 is abrupt and henceforth synthetic variants decline gradually or disappear altogether (as for *time wherein*), while the analytic variants gain in proportion.

The shift away from synthetic and toward analytic renderings is not only in line with the "typological drift [of English] from synthetic to analytic" mentioned by Nevalainen & Raumolin-Brunberg (2012, p. 203, referring to Sapir 1921, p. 165–168)

but by creating highly conventionalized, formulaic patterns it also seems to strongly contribute to syntagmatic predictability of RCs in scientific English.

7.3.2.2 German

Unlike in English, the most frequent lexical trigrams preceding *welch.** in scientific German (Table 7.7) do not function as manner adverbials or quantifiers, but they are mostly demonstrative pronouns further specified by the following RC. Also, the lexical trigrams do not coincide with the most frequent grammatical contexts, i.e. *adjective noun comma*. Instead, in all but the last period, we see a frequent occurrence of demonstrative pronouns (*diejenigen,*) functioning as the syntactic head of the RC. Moreover, the trigrams reflect the shift in punctuation from the historical virgule (*die jenige /*) to the standardized usage of a comma preceding RCs (*daß diejenigen,*). We can furthermore note that the demonstrative pronoun was formerly written as two words and seems to have become standardized as a compound demonstrative between 1700 and 1749.

While in 1700, demonstrative pronouns seem to occur in subordinate complement clauses introduced by the complementizer *daß* (Example (10)), in 1750 and 1800, they predominantly occur in sentence-initial position (Example (11)).

(10) *Man hat observirt,* ***daß diejenigen welche*** *am scharbock laboriren, offt von dem safft der Brunn-Kresse curiret worden.* (Hans F. von Flemming, Der Vollkommene Teutsche Jäger, 1724)

(11) ***Diejenigen, welche*** *durch Verblutungen, oder heftige speichelkuren, fast alle ihre säfte eingebüßt haben, und also gleichsam von neugeschaffnen säften leben, erlangen ihr altes Temperament wieder.* (Albrecht von Haller, Anfangsgründe der Phisiologie des menschlichen Körpers, 1762)

The constructions point to the fact that in scientific German there seems to be a high usage of pronominal demonstrative reference until the first half of the 19$^{\text{th}}$ c. In terms of the syntagmatic predictability of RCs, it can be observed that these structures exhibit a high degree of predictability. This is due to the fact that the demonstrative pronoun *d(er/ie/as)jenige(n)* requires additional specification, which is most commonly achieved through the use of an RC, and less frequently through the employment of a prepositional phrase, as exemplified in Example (12).

(12) *Die Arten von Corophium graben sich Löcher in den Schlamm,* ***diejenigen von Cerapus*** *bauen sich, wie die Larven der Phryganiden, cylindrische Gehäuse, welche sie mit sich schleppen.* (Alfred E. Brehm, Illustrirtes Thierleben. Bd. 6., 1869.)

Examples (10) and (11) show that the construction is a highly explicit rendering of concepts lacking a technical term. For instance, *diejnigen, welche am Scharbock*

period	Freq pM	Freq raw	trigram	semantic function	syntactic function
1650	3528.29	52	die jenige /	demonstrative pronoun	-
	1899.85	28	die jenigen /	demonstrative pronoun	-
	1832.00	27	der Linie /	NP	-
1700	1413.27	38	Tochter, mit	-	NP + prep. object /attribute
	1264.50	34	daß diejenigen,	demonstrative pronoun	-
	1115.74	30	zu sehen,	-	to-infinitive
1750	2644.88	157	. Diejenigen,	demonstrative pronoun	-
	1212.94	72	zu sein,	epistemic	to-infinitive
	1179.25	70	als die,	pronoun	-
1800	2967.13	157	. Diejenigen,	demonstrative pronoun	-
	1360.72	72	zu sein,	epistemic	to-infinitive
	1322.93	70	als die,	pronoun	-
1850	1567.25	194	ist es,	HN topicalization	cleft
	1147.16	142	Zeit, in	temporal	NP + prep. object /attribute
	1009.82	125	sind es,	HN topicalization	cleft

Table 7.7: Lexical trigram context of *welch.** in scientific German (DTAW).

laborieren (those who suffer from scurvy) could be rephrased by the compound *Skorbutkranke* (sufferers of scurvy). The disappearance of the trigram in 1850 could be explained by an increasing number of technical terms.

Between 1700 and 1800, *zu*-infinitives become another highly frequent lexical context of *welch.**. The *zu*-infinitives belong to verbs expressing epistemic modality such as *scheinen* (Example (13)).

(13) *Dieses* ***scheinet*** *eben diejenige Materie* ***zu sein, welche*** *nicht selten, das Fleisch leuchtend macht, und welche in Gestalt der Flammen aus todten Körpern, wie auch aus Thieren, gefahren sein soll.* (Albrecht von Haller, Anfangsgründe der Phisiologie des menschlichen Körpers, 1762)

In 1850, constructions, such as *ist es,* or *zu sein,* become strongly favored. The verbal fragments belong to cleft constructions with topicalized noun phrases functioning as the antecedents of the relativizer *welch.** (Examples (14) and (15)). The constructions

reflect a pragmatic trend in scientific German texts to mark the focus on the topic in a sentence.

(14) *Gerade **diese Zugehörigkeit** zu einer und derselben Gruppe von Vorstellungen **ist es, welche** hier die Annahme von Ähnlichkeitsassoziationen rechtfertigt.* (Emil Kraeplin, Ueber die Beeinflussung einfacher psychischer Vorgänge durch einige Arzneimittel, 1892)

(15) *Vier Momente **sind es, welche** diese ungünstige Veränderung zur Folge haben.* (Josef von Lehnert, Die Seehäfen des Weltverkehrs, 1891)

In general German (Table 7.8), the most frequent lexical trigrams in 1650 are very similar to those in scientific German (*die jenige /*, *den jenigen /*). Also, the sentence-initial demonstrative pronoun (*. Diejenigen,*) is amongst the most frequent contexts in 1750.

General German, in contrast to scientific German, shows a high occurrence of actual nominal RC heads (*Wurzel, Fluß, Zeit, Tage*). A look into the corpus reveals that the 29 occurrences of *Wurzel/ aus* stem from the same text (i.e. Zwinger, 1690; Example (16)); also, the 173 occurrences of *, Fluß, welcher* and the 47 occurrences of *in England,* both stem from the same source (Hübner, 1704; Examples (17) and (18)).

(16) *Der Goldapffel hat ein zertheilte **wurtzel / auß welcher** sehr lange / schwache / inwendig hole / langhaarige / zur erden sich neigende / ästichte stengel wachsen / an welchen die blätter hangen / etwas breit / groß / tieff zerkerfft / bleichgrün / und eines starcken unfreundlichen Geruchs.* (Theodor Zwinger (der Jüngere), Das ist Theatrvm Botanicvm: Neu Vollkommenes Kräuter-Buch, 1690)

(17) *Trebia**, Fluß, welcher** im Genuesischen Gebiet entspringet, und sich oberhalb Piacenza in den Po ergeust.* (Johann Hübner, Reales staats- und Zeitungs-Lexicon, 1704)

(18) *Kent, Cantium, Provintz **in Engelland, welche** gegen Westen an Essex, surrey und sussex grentzet, gegen Osten aber von dem Meer umgeben, und von Franckreich durch den Pas de Calais abgesondert wird.* (Johann Hübner, Reales staats- und Zeitungs-Lexicon, 1704)

On the other hand, the occurrences of *Zeit, in* and *Tage, an* stem from various different sources (Examples (19) and (20)).

(19) *Wie wäre es sonst auch zu erklären, daß in einer **Zeit, in welcher** die Beschäftigung mit geistigen Dingen so allgemein ist, das Glück so selten gefunden wird?* (NA, Unsere moderne Bildung im Bunde mit der Anarchie, 1852)

period	Freq pM	Freq raw	trigram	semantic function	syntactic function
1650	2453.81	51	die jenige /	demonstrative pronoun	-
	1395.30	29	Wurzel / aus	-	NP + prep. object /attribute
	1347.19	28	den jenigen /	demonstrative pronoun	-
1700	4684.03	173	, Fluß,	NP	
	1868.20	69	von denen,	demonstrative pronoun	-
	1272.54	47	in England,	PP	-
1750	2807.97	62	zu machen,	-	to-infinitive
	1856.88	41	. Diejenigen,	demonstrative pronoun	-
	1585.14	35	als die,	comparative	-
1800	1625.54	33	als die,	comparative	-
	1428.50	29	Zeit, in	temporal	NP + prep. object /attribute
	1083.69	22	ist, in	-	NP + prep. object /attribute
1850	1661.13	42	Zeit, in	temporal	NP + prep. object /attribute
	1067.87	27	ist es,	HN topicalization	cleft
	909.67	23	Tage, an	temporal	NP + prep. object /attribute

Table 7.8: Lexical trigram context of *welch.** in general German (DTAG).

(20) *Aber an dem **Tage, an welchem** auch der Papa Quakatz hinter mir zum erstenmal fragte: Wie war die Geschichte, Junge?* (Wilhelm Raabe, Stopfkuchen, 1891)

Our general German corpus consists of texts classified as "Gebrauchsliteratur", i.e. narrative non-fiction, and "Belletristik", i.e. narrative fiction. The text sources of the first three trigrams belong to the sub-class narrative non-fiction, more precisely encyclopedic works, which is a plausible explanation for the accumulated occurrence of the same lexical pattern since encyclopedic entries are usually structured in a standardized way. In 1850, cleft constructions also appear among the three top lexical trigrams in general German. They, too, stem from the non-fictional part of the general German corpus. We can conclude from these findings that within general German, the most conventionalized lexical patterns introducing RCs are used in non-fictional

texts for special purposes, which is not surprising since special-purpose literature can be assumed to show more standardized structures than narrative fiction.

In summary, *welch.** shares similar lexical contexts in both scientific and general German. The qualitative analysis of the most frequent patterns in general German has shown that especially the non-fictional portion of the general German corpus shows a preference for conventionalized patterns in which RCs tend to occur. Having said that, we can conclude that highly predictable patterns introducing RCs with the relativizer *welch.** are more characteristic of specialized and scientific German literature than of narrative fiction. Regarding the form of the lexical patterns in general German compared to scientific German, we found that scientific German shows a preference for contexts consisting of function words rather than content words. RCs in scientific German initially mostly define demonstrative pronouns creating pronominal reference. In the 18th c. head nouns are introduced in epistemic modality using the *zu*-infinitive construction (*scheinen* + *zu sein*) and in the 19th c., RCs occur increasingly in sentences with topicalized head nouns as parts of cleft constructions. The lexical patterns do not reflect the most frequent POS trigrams representing premodified noun phrases of the form *adjective noun comma*. The fact that these are not reflected in the most frequent lexical trigrams indicates that these patterns do not represent conventionalized lexical patterns. Instead, the patterns that we found represent typical lexico-grammatical syntagmatic contexts of RCs in scientific German. In general German, apart from the shared preference for pronominal antecedents, we found a distinctive preference for the use of temporal expressions, such as *Zeit, in* .

7.4 Summary

In the first part of this chapter (Section 7.1), we have shown that RCs in scientific writing become increasingly predictable. In scientific writing compared to general writing, RCs show a comparatively stronger trend toward lower surprisal, i.e. higher predictability. In Section 7.2, we found that the overall lower surprisal of RCs in scientific writing in both English and German can be attributed to the decreasing surprisal of the preferred relativizers *which* and *welch.**, which have become increasingly predictable over time. In Section 7.3, we conducted a qualitative analysis of the three most frequent grammatical (POS trigrams) and lexical (lexical trigram) contexts. We found that both in English and in German scientific writing there is a clear trend toward the preference for specific grammatical contexts. General writing in both languages proves to be less targeted at one specific preferred grammatical context. In scientific English, the preferred grammatical context of *which* is the POS trigram *determiner noun preposition*, while in scientific German the preferred grammatical contexts of RCs are complex noun phrases with adjectival premodification (*adjective noun comma*). The lexical contexts in German and English have shown that between registers there is a great overlap in most frequent trigrams preceding *welch.** and

which. The lexical syntagmatic contexts of RCs in scientific English coincide with the grammatical patterns and represent expressions of manner (*the manner in*) and quantification (*, one of*). While the lexical contexts are chiefly shared between general and scientific English, we find a slight preference for quantification patterns in scientific English and a slight preference for manner expressions in general English. In German, the lexical contexts of *welch.** do not coincide with the preferred grammatical patterns. As in English, many of the lexical trigrams are shared between the two registers as well, especially demonstrative pronouns. Apart from these, the two German registers differ in that general German shows higher usage of content words functioning as head nouns of the RCs expressing temporal relations such as *Zeit, in*, which apart from occurring frequently, also show a dispersion across 17 different texts in 1850. In scientific German, the lexical contexts of *welch.** at the end of the observed time span mostly consist of function words such as *ist es,* syntactically forming fragments of cleft constructions and semantically functioning as parts of focus clauses with a topicalized head noun. Overall, we can confirm our hypothesis that scientific writing has developed toward higher predictability of RCs (H1.2a) due to the increasingly preferred lexical choice of the register-specific relativizer and its increasingly conventionalized usage in context (H1.2b). Moreover, the analysis of grammatical contexts (POS trigrams) has shown that scientific German compared to scientific English develops later toward lower grammatical complexity (H2).

Chapter 8

Summary of Part III

In this first part of our analyses, we have investigated the development of lexico-grammatical complexity in scientific English and German. Our hypotheses were that scientific writing should become less complex in terms of paradigmatic richness of the relativizer paradigm (H1.1), i.e. show lower *entropy* (uncertainty about the choice of the relativizer) over time, and that scientific writing should develop towards higher syntagmatic predictability in terms of lower *surprisal* of RCs when they are used. Both measures reflect a trend towards a conventionalized usage of lexico-grammatical structures, which we can expect to emerge during register formation. Cognitively, the measures reflect expectation-based processing effort and they can show us how scientific writing has evolved towards a conventionalized, more efficient code for expert readers over time. Our results for paradigmatic richness (Section 6.2) have shown that, indeed, the relativizer paradigm in scientific writing decreases in entropy over time. The paradigms in both languages have shifted towards a preference for certain relativizers (*which* and *welch.**) and a disfavor of other options (Section 6.3). The choice of one preferred option and the abandonment of alternative options helps a reader to expect the upcoming word and reduce processing effort in the moment of encountering the word. By converging on a preferred option, scientific writing has become more efficient and lexico-grammatically less complex. The comparison with the development of general language has shown that in English the trend toward lower paradigmatic richness is specific to scientific writing due to a distinctive preference for *which*, while entropy in general English did not change markedly over time. In German, we found that general language actually seemed to drive the development towards lower paradigmatic richness (due to a clear preference for *d.**) and reduced entropy much earlier than scientific German, which preserved its paradigmatic richness up until the middle of the 19$^{\text{th}}$ c.

Our second study on syntagmatic predictability (Chapter 7) has shown that overall RCs in scientific English and German become more predictable, i.e. less surprising (Section 7.1). Compared to general language, we found that in English the higher

predictability is again specific to scientific writing, while in German the surprisal of RCs decreases across registers. In line with the finding that scientific writing over time converges on one preferred relativizer, our surprisal analysis of specific relativizers has shown that compared to the other relativizers, *which* and *welch.** become the most predictable relativizers over time in scientific English and German given their syntagmatic contexts (Section 7.2). The qualitative analyses of the syntagmatic grammatical (Section 7.3.1) and lexical (Section 7.3.2) contexts have shown that in scientific language RCs introduced by the relativizers *which* and *welch.** increasingly occur in strongly conventionalized contexts conveying quantification and manner in English and in German as forming part of cleft constructions creating focus sentences.

Overall, we found fewer differences in the trends across registers in German and more register-specific trends in English. This is not extremely surprising and can be explained by the fact that scientific writing in the German vernacular started to develop much later than in English. Another factor is that in German the registers are less clear-cut than in English. The scientific English corpus includes texts from the same publisher and only contains natural-scientific texts, while the German scientific corpus consists of texts from various different publishers and includes texts from an enormously diverse set of topics. Moreover, the general German corpus contains narrative-fictional texts as well as non-fictional texts for specific purposes. The latter can also be assumed to become strongly conventionalized over time, which might explain the similar trends found for scientific and general German.

Part IV

Corpus Studies: Syntactic Complexity

Chapter 9

Syntactic Intricacy

Having analyzed the development of lexico-grammatical complexity in English and German over time, we will now shift our focus to syntactic complexity. We start by analyzing *syntactic intricacy* created by relative clauses (RCs). Halliday & Webster (2004, p. 33) describe grammatical intricacy as "the length and depth of the tactic structures whereby clauses come together to make up a clause complex". Thus, we believe that the usage of RCs as grammatically highly explicit material represents one way of modulating the syntactic intricacy of a sentence. In the present analysis, we measure intricacy in terms of the relative frequencies of RCs in the corpora at hand. For this analysis, we use the Universal Dependencies UD-parsed version of the corpora (described in Section 4.3) and we count RCs as the number of occurrences of the UD-relation `acl:relcl`. As formulated in hypothesis H1.3a, we expect that RCs overall will become less frequent in scientific writing, since they represent explicit ways of noun phrase post-modification especially useful to explain formerly unknown (or undefined) concepts. Over time, we assume that the need for explicit definitions in scientific discourse decreases since the scientific communities increasingly rely on ample shared knowledge and terminology, making explicit renderings superfluous. While H1.3a focuses on the overall syntactic intricacy in scientific writing, in hypothesis H1.3b, we set forth the assumption that RCs will not only become less frequent overall, but also the number of RCs per sentence will decrease, since multiple embeddings within a single sentence substantially increase the sentence-internal intricacy and thus make a sentence especially hard to process.

To test our assumptions, we use the number of sentences per 50-year period as the normalization base, i.e. the frequency of RCs per 1000 sentences (Section 9.1). The underlying assumption here is that a higher number of RCs per 1000 sentences represents a stronger preference for explicit noun phrase post-modification, which would not be captured for instance by a normalization per million tokens, since we would not detect whether RCs occur in many short or few long sentences. To verify this, we also inspect normalized frequencies per 1 million tokens for comparison and

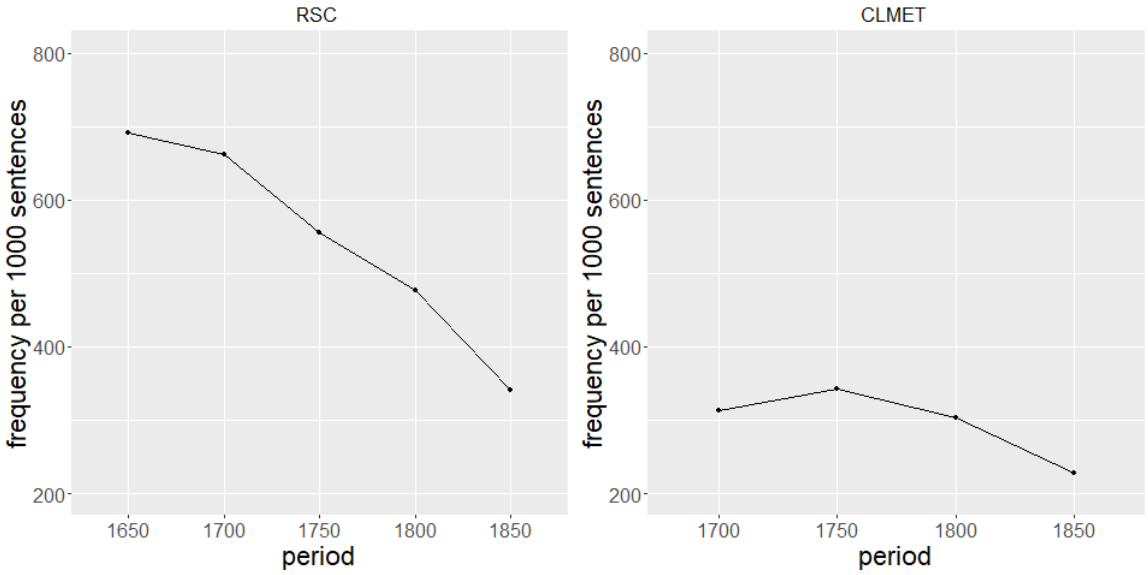
use the average sentence length to interpret our results in more detail.

Secondly, we analyze the degree of embeddedness of sentences that include at least one RC by calculating the average number of RCs within one sentence including at least one RC (Section 9.2). The measure of embeddedness (the average number of RCs per sentence per 50 years) is even more specific since it is calculated on the basis of the number of sentences that *do* include at least one RC. Through this, it is possible to show the actual sentence-internal intricacy created by RCs in a sentence.

9.1 Relative frequencies of RCs

9.1.1 English

We start by normalizing the number of RCs by 1000 sentences to see how frequent RCs are in relation to the number of sentences in each corpus. This basis of normalization is tightly connected to the development of sentence length over time, i.e. if sentences are longer, there are more tokens per 1000 sentences than when sentences are shorter. Overall, we find that RCs are much more frequent among 1000 sentences in scientific English (Figure 9.1a) than in general English (Figure 9.1b).



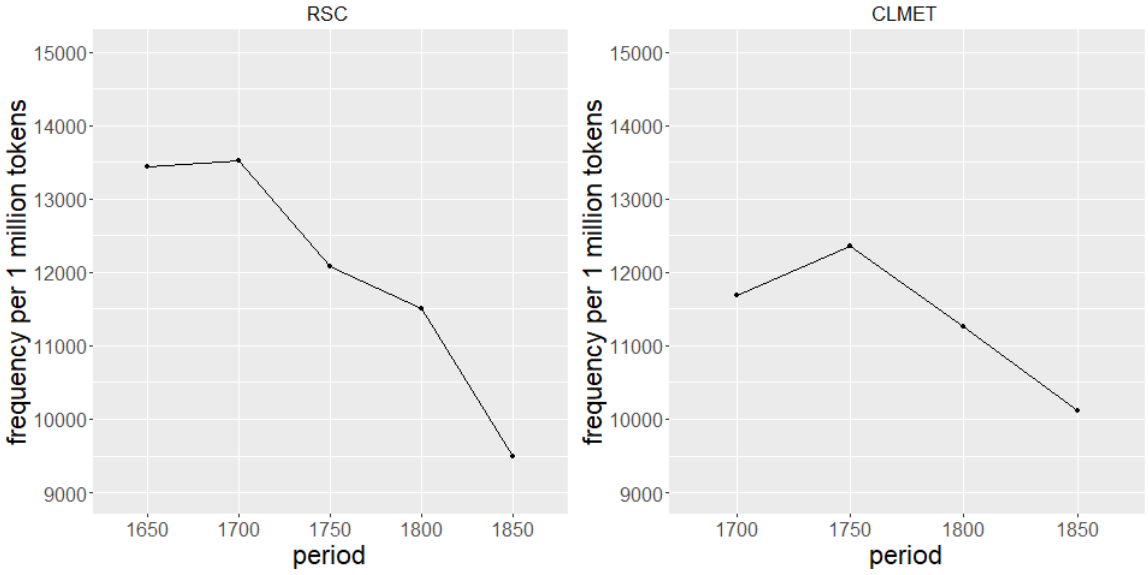
(a) RSC **(b)** CLMET

Figure 9.1: Frequencies of RCs per 1000 sentences in (a) scientific (RSC) and (b) general (CLMET) English by 50-year periods.

Moreover, the trends in the two corpora differ remarkably. In 1650, scientific English exhibited a frequency of nearly 700 RCs per 1000 sentences; this was halved by 1850. In contrast, general English contained approximately 300 RCs per 1000 sentences in 1700 and just under 200 by 1850. These findings show that RCs are

less prevalent in general English compared to scientific English and their decrease in frequency over time is much slower.

For comparison, we calculate RC frequencies normalized by 1 million tokens (Figure 9.2). We find that this type of normalization makes the differences between the corpora seem smaller. This is possibly due to the fact that sentences are shorter in general English. In other words, if the corpus has many short sentences (with few tokens), RCs are distributed across more sentences and appear less frequent than when measured across tokens, which have comparatively lower frequencies than a corpus with many long sentences. The corpus statistics of the parsed corpora in Section 4.3.3 corroborate this explanation.



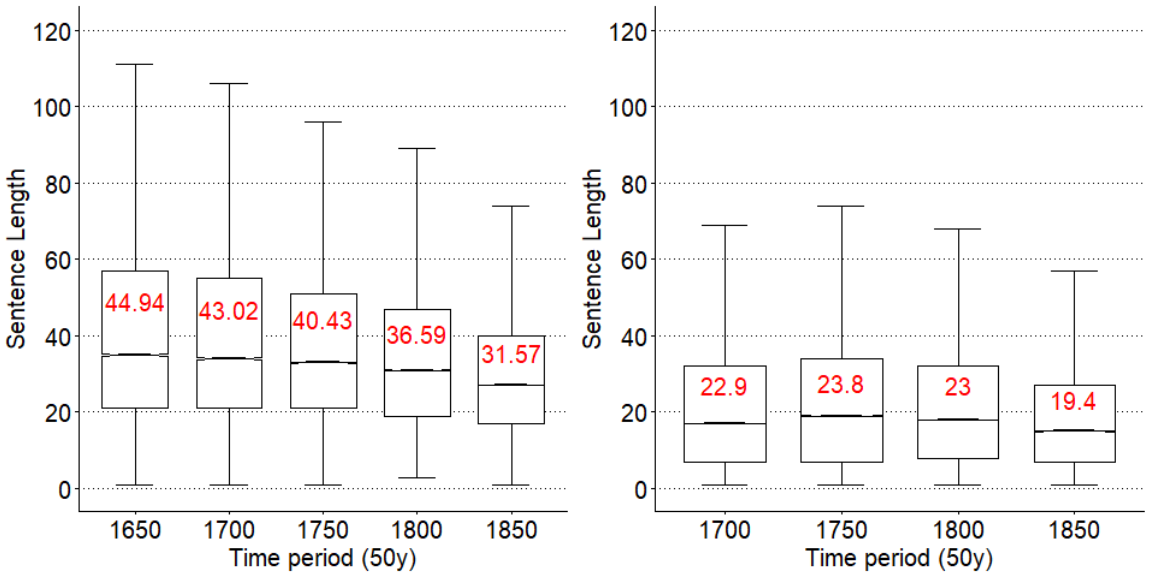
(a) RSC **(b)** CLMET

Figure 9.2: Frequencies of RCs per **1 million tokens** in (a) scientific (RSC) and (b) general (CLMET) English by 50-year periods.

We can derive from this that general English has overall shorter sentences (and also a decreasing trend), leading to a higher proportion of RCs per 1 million tokens and a lower proportion per 1000 sentences. Figure 9.3 corroborates this assumption: In general English, sentences are overall much shorter than in scientific English. For this reason, RCs are less frequent amongst the high number of sentences. Since the sentences themselves are short, the number of tokens that RCs occur in is relatively low, which makes RCs look more frequent when normalized per 1000 sentences.

In summary, the analysis has shown that RCs become less frequent in both corpora. Scientific English undergoes a more drastic change than general English, showing initially higher RC frequencies and a linear and steep reduction of RC usage over time, while general English shows an overall lower RC usage and a less pronounced decline of RCs. We can learn from these insights that syntactic intricacy used to be a typical feature of scientific English writing, gradually turning in the opposite

direction by drastically reducing the intricacy created by RCs.



(a) RSC **(b)** CLMET

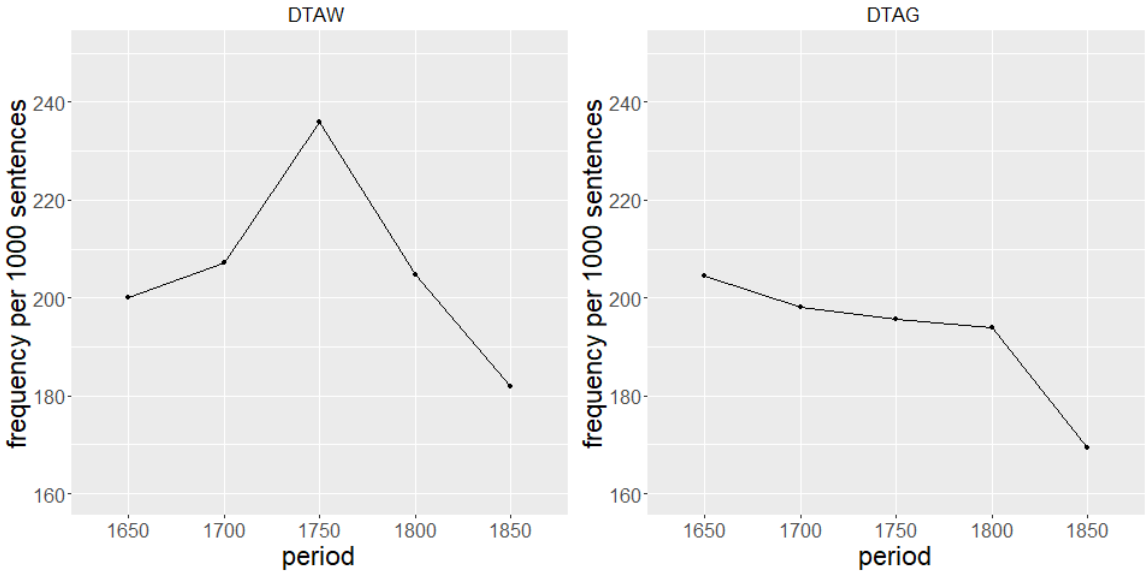
Figure 9.3: Distribution of sentence lengths per 50-year period in (a) scientific (RSC) and (b) general (CLMET) English.

9.1.2 German

Examining the normalized frequencies of RCs per 1000 sentences in the two German corpora, divergent trends emerge. Specifically, as anticipated, in scientific German (Figure 9.4a), RCs initially become more frequent until the end of the 18th c. and subsequently decrease even more sharply. In general German (Figure 9.4b), however, the trend is consistently downward. These findings align with our Hypothesis H2, which posits that scientific syntax initially tends towards greater complexity before subsequently decreasing in complexity in the late 19th c.

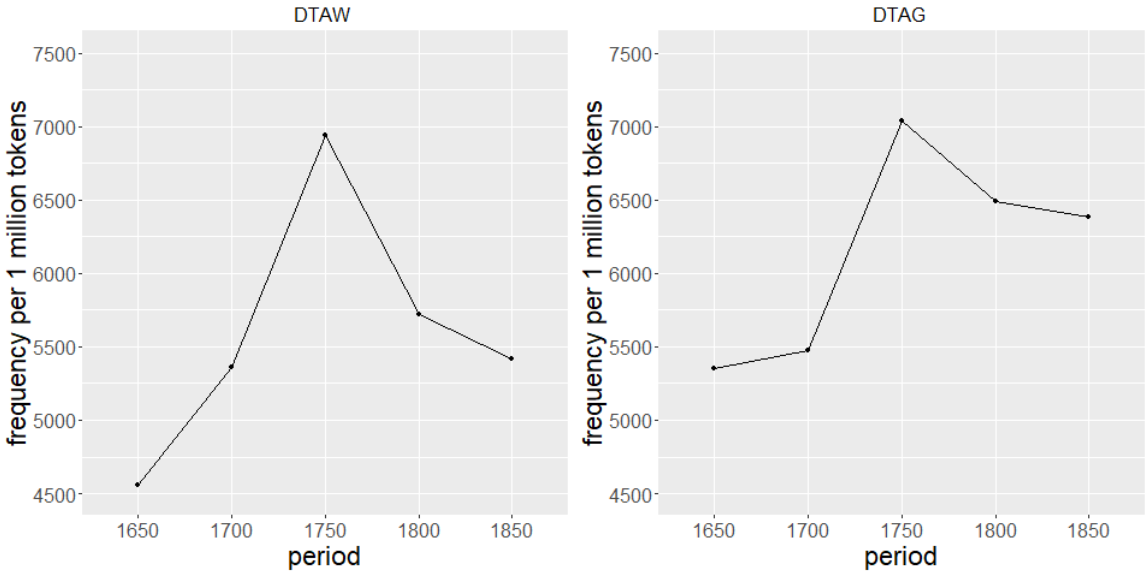
Comparing this base of normalization to the trend for RCs per 1 million tokens (Figure 9.5), we see a stunning difference between the trends in general German (Figure 9.5b): While frequencies per 1 million tokens increase over time, RC frequencies normalized by 1000 sentences decrease.

As done for English, inspecting the corpus sizes in terms of tokens and sentences (Section 4.3.3, Tables 4.7 and 4.8 and Figures 4.6 and 4.7) reveals the underlying causes for the striking discrepancies of RC frequencies at different normalization bases. In general German, the simultaneous upward trend per 1 million tokens and the downward trend of RCs per 1000 sentences can be explained by the diverging trends in total sentences and tokens per time period. In general German, the corpus size in terms of tokens indeed decreases (Figure 4.7a) while in terms of sentences, it increases (Figure 4.7b).



(a) DTAW **(b)** DTAG

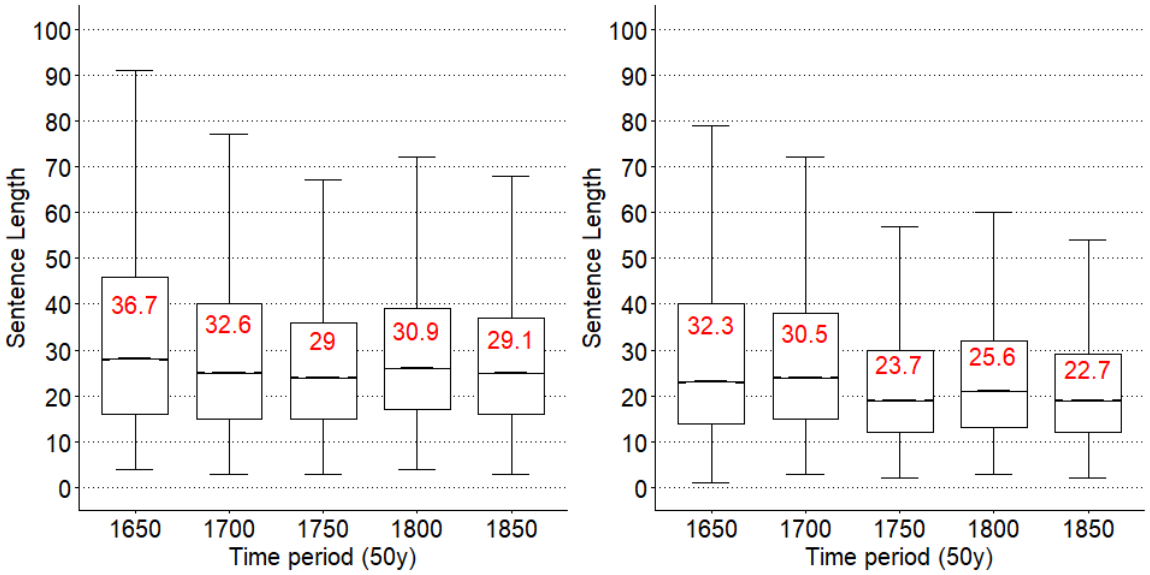
Figure 9.4: Frequencies of RCs per 1 thousand sentences in (a) scientific (DTAW) and (b) general (DTAG) German by 50-year periods.



(a) DTAW **(b)** DTAG

Figure 9.5: Frequencies of RCs per 1 million tokens in (a) scientific (DTAW) and (b) general (DTAG) German by 50-year periods.

In line with this observation, the decrease in sentence lengths in general German (Figure 9.6b) contributes to explaining the diverging relative frequencies: the sentences in general German become shorter, and thus, per token, RCs become more frequent, since RCs occur in increasingly short sentences. For scientific German, the development of sentence length over time shows that RCs first increase in frequency



(a) DTAW **(b)** DTAG

Figure 9.6: Distribution of sentence lengths per 50-year period in (a) scientific (DTAW) and (b) general (DTAG) German.

independently of sentence length. The peak of RC usage can be observed in the period of 1750, which also happens to be the period with the lowest mean sentence length. This means that even if sentences become shorter over time, the usage of RCs is not affected by this.

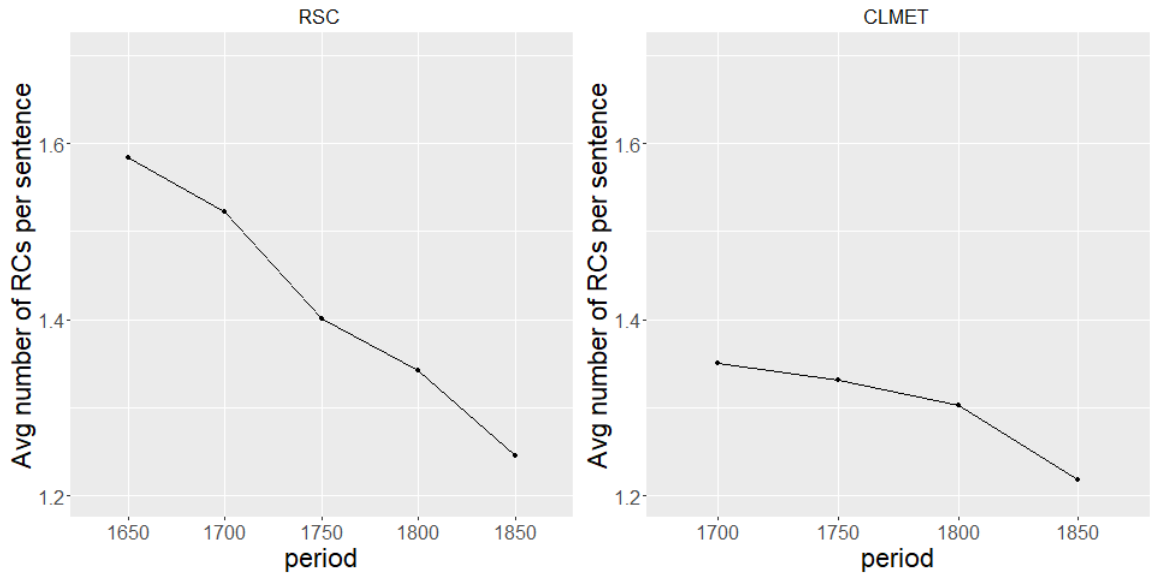
9.2 Number of relative clause embeddings per sentence

9.2.1 English

As a second regulator of syntactic intricacy, we now analyze how many RCs on average are embedded within one sentence; in other words, if a sentence has at least one RC, how many RCs does it contain on average?

For this, we extract all sentences with at least one RC and divide the total number of RCs per 50 years by the number of sentences including one or more RCs to get the average number of RC embeddings per sentence. We find that in both English corpora (Figure 9.7), the average number of RCs per sentence decreases over time. The scientific corpus (Figure 9.7a) shows a straight and steep downward trend from an initial 1.58 RCs per sentence down to 1.25 RCs per sentence. The high average reflects the initially frequent use of several RC embeddings per sentence as in Example (1).

General English (Figure 9.7b) starts out with on average fewer RC embeddings per sentence than scientific English in 1700 (1.35) and drops to 1.22 in 1850. Our



(a) RSC **(b)** CLMET

Figure 9.7: Average number of RCs per sentence in (a) scientific (RSC) and (b) general (CLMET) English by 50-year periods.

findings confirm our hypothesis H1.3b that scientific English develops toward lower embeddedness of RCs per sentence. The steep and almost linear trend observed in scientific English suggests that the use of strong RC embeddedness is gradually declining, indicating a shift in the preference for this syntactic feature.

(1) *"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)

In summary, our analyses on syntactic intricacy have confirmed our hypotheses that in scientific English writing, RCs overall become less frequent as normalized by the total number of sentences as well as tokens of each time period (H1.3a). Secondly, the number of RCs per sentence also decreases almost linearly (H1.3b).

The similar trends in the calculated measures are not surprising due to a plausible correlation between them. However, the described developments in scientific English are stunningly linear in their trend toward lower intricacy. Compared to general English, scientific English starts out at an elevated level of intricacy, but ends up at a similar level of intricacy created by RCs as general English. Moreover, compared

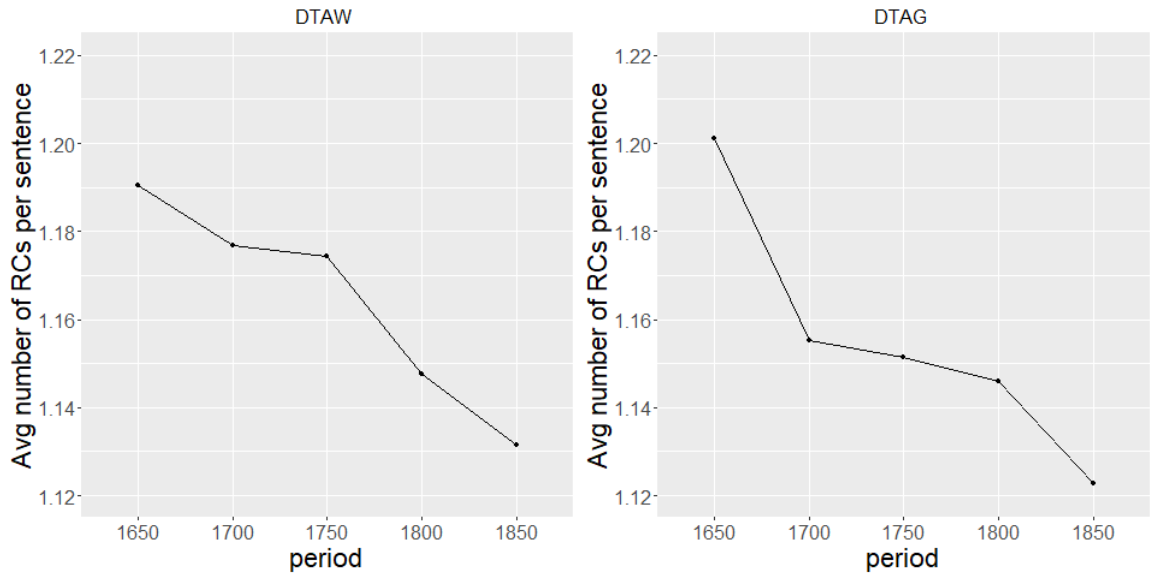
to relatively stable sentence lengths in scientific English, the decrease in syntactic intricacy is remarkable, meaning that although sentences stay relatively long, their intricacy goes down. This is different in general English, where RC frequency seems to be a result of the general turn towards building shorter sentences. In summary, scientific writers seem to make ample use of RCs initially, when they are necessary to express explicit descriptions of formerly unknown concepts. Over time, RCs are used with decreasing frequency. General English, in contrast, throughout all time periods makes less use of RCs than scientific English while also showing a decreasing tendency, albeit one that is less pronounced than that in scientific writing. In terms of register formation (H1), we can clearly see that RCs as a means of creating syntactic intricacy become increasingly disfavored.

9.2.2 German

Calculating the average number of RC embeddings per sentence in the German corpora shows that the trends in both scientific (Figure 9.8a) and general (Figure 9.8b) German are surprisingly similar, both showing a downward trend in average RC embeddings per sentence. The average embeddings range between a maximum of 1.2 and a minimum of 1.12 RCs per sentence. The average is comparatively low compared to the English average number of RC embeddings, ranging between a maximum of 1.58 and a minimum of 1.2 RCs per sentence. In German, the use of accumulated RCs in one sentence seems to become disfavored across both registers. It is, however, in line with our hypothesis that scientific German decreases in syntactic complexity later than general German does, since the average number of RCs decreases most sharply after 1750, while stabilizing between 1700 and the period of 1750. In general German, on the other hand, the sharpest decrease happens much earlier, i.e., between 1650 and the period of 1700.

In summary, the results obtained for syntactic intricacy in German have shown that the relative frequencies of RCs, both normalized by tokens and sentences, show a climactic shape with a peak at the end of the 18^{th} c. Furthermore, the analysis of the average number of RCs per sentence has revealed a gradual decrease over time, with a relatively mild decline in the first three periods followed by a more pronounced decrease in the last two periods. The results confirm both our hypotheses H1.3a and H1.3b, assuming that RCs in scientific writing become less frequent over time, as well as our hypothesis H2 that German scientific writing shows a time-shifted development toward lower grammatical complexity.

We may add to these conclusions that the trends for scientific and general German differ remarkably, where RCs increase when measured against the total token size of the corpus and decrease when measured against the sentence size of the corpus. The obtained trends reflect the fact that in scientific German, the corpus size in tokens seems to be positively correlated with the corpus size in sentences, effectively leading to a relatively stable sentence length. Keeping this in mind, the peak in



(a) DTAW **(b)** DTAG

Figure 9.8: Average number of RCs per sentence in (a) scientific (DTAW) and (b) general (DTAG) German by 50-year periods.

RC frequency at the end of the 18th c. is independent of sentence length and points to an independent preference for RCs as means of noun phrase modification during this time. The climactic trend observed in scientific German suggests that RCs were first discovered as a useful means to describe unknown matters, but were eventually abandoned as other less explicit renderings became sufficient due to increased shared expert knowledge among scientific communities.

9.3 Summary and interpretation of results

Our analysis of syntactic intricacy was based on the hypothesis that RCs as means of syntactic intricacy would decrease in scientific writing (H1.3). The analysis has shown that broadly speaking, there is a cross-linguistic and cross-register trend towards lower syntactic intricacy, since the relative frequencies of RCs as well as the number of RC embeddings per sentence have been shown to decrease over time. English and German, however, differ in that German shows a climactic development, first increasing in intricacy and then decreasing, which is in line with the trend indicated by previous work (Habermann, 2011; Admoni, 1990) on scientific German first becoming more intricate with stronger use of hypotactic structures, and then showing a trend of disentanglement afterward. The contrast between English and German shows that scientific English started with high syntactic intricacy and gradually abandoned it, while German only began to discover clausal subordination as a useful means of noun phrase modification and increasingly made use of it. The eventual turn away from this high intricacy in scientific German is in line with our hypothesis that scientific

writing started to disentangle its intricate syntax from the beginning of the 19th c. onward, which is also based on observations made by Möslein (1974) and Admoni (1972, 1990).

In terms of register differences, our data confirm the claims made in previous work (e.g. Biber & Gray, 2011b, 2016) that in scientific English, RCs are overall more frequent than in general (non-scientific) English. Our data also reflected the suggested development of sentence length, which in previous work was described as becoming shorter in both languages and registers. In Part IV, we found that the lexico-grammatical variability of RCs, i.e. the paradigmatic richness in the relativizer paradigm, decreases and the syntagmatic contexts of RCs become more conventionalized and thus predictive of RCs to occur. The two findings combined, i.e. the conventionalization of the RCs that we do find and the fact that RCs become overall less frequent, suggest that two processes of language change seem to be reflected in our findings: RCs seem to undergo a process of "preference in performance", as well as a "Miniminze Domain" process (Hawkins, 2004). The two principles state that

1. "Grammars have conventionalized syntactic structures in proportion to their degree of preference in performance, as evidenced by patterns of selection in corpora and by ease of processing in psycholinguistic experiments" (Hawkins, 2004, p. 3).

2. According to the principle of "Minimize Domains", the human language processor tends to favor the use of the smallest syntactic domains possible when assessing a particular grammatical relationship. Similarly, the principle of "Minimize Forms" suggests that reducing the formal complexity of each linguistic form is preferred (cf. Hawkins, 2004).

By reducing the number of comparatively long syntactic domains such as RCs and at the same time converging on a strongly conventionalized usage, scientific writing seems to reflect both principles. We can assume that the partial abandonment of RCs overall is due to their replacement by shorter renderings such as non-clausal noun phrase modifications (attributive adjectives, post-modifying prepositional phrases, etc.) or even the creation of new terminology combining semantic content of concepts that formerly had to be described explicitly before becoming so well-known that a name for them could be agreed upon. Apart from the speculations about what might have happened to the RCs that were abandoned over time, in Part IV we got a glimpse into what the "surviving" RCs look like. The remaining RCs increasingly use one preferred relativizer (i.e. *which* or *welch.**) and occur in increasingly similar grammatical contexts. Among RCs introduced by *which*, especially partitive constructions, which are indeed not possible to express in a non-clausal construction (Example(2)), become increasingly frequent.

(2) a. *I ate five apples one of which was rotten.*
b. **I ate five including one rotten apple.*

c. *I ate five apples* ***and*** *one of them was rotten.*

However, English also shows high usage of RCs representing adverbial clauses of manner, which *are* possible to replace in shorter terms:

(3) a. *A short description is given of* ***the manner in which*** *the observations were made.*
b. *A short description is given of* ***how*** *the observations were made.*

In German, we see increasing use of RCs following focus sentences using cleft constructions for topicalization of the head noun.

(4) *Vier Momente sind es, welche diese ungünstige Veränderung zur Folge haben.*
(Josef von Lehnert, Die Seehäfen des Weltverkehrs, 1891)

In this kind of construction, it is not the denotative meaning of the sentence that is affected by using the cleft construction, but rather the pragmatic motivation of focusing on the head noun. In such cases, even if a shorter rendering would grammatically be possible, the sentence would not convey the same meaning. We can therefore assume that these constructions endure due to their pragmatic value and the fact that they cannot be substituted by shorter renderings without a pragmatic loss.

Our second hypothesis related to syntactic intricacy was that on average there will be fewer clausal RC embeddings per sentence (H1.4). The more RCs a sentence contains, the more complex and harder to process the sentence becomes. Regarding the average number of RC embeddings, we made two interesting observations. First, we saw that in both languages, the accumulated use of RCs within a sentence seems to become disfavored and the average number of RCs per sentence declines. The trends are almost linear in both languages and in both registers. Second, we found that in English RC embeddedness was initially much stronger than in the German corpora. We can thus conclude that both languages show trends of decreasing syntactic intricacy: English from the beginning onward, and German in a time-shifted manner starting in the 19$^{\text{th}}$ c.

Chapter 10

Locality

Having analyzed the development of relative frequencies of relative clauses (RCs) in scientific language to trace syntactic *intricacy* in Chapter 9, in the present chapter, we turn to investigating syntactic complexity from the viewpoint of dependency locality as calculated by dependency length (DL). DL describes the distance in tokens between a syntactic head and its dependent and ADL is the average length of all dependencies within a sentence (described in detail in Section 5.2.2). As we have seen in the last chapter, in terms of frequency, scientific language loses part of its syntactic complexity by relying less on the syntactic embedding of RCs. The present chapter aims to trace the implications that the decrease in RCs has on average dependency length (ADL) influencing processing effort by modulating working memory demand. The hypothesis (H1.4a, Section 3.1.2.2) here is that scientific language over time is characterized by increasing locality, i.e. shorter dependency length (DL).

The chapter is structured as follows: We start by conducting a macro-analysis to answer the question of whether ADL decreases diachronically in both languages and across registers (Section 10.1). To do so, we inspect ADL in two different ways. First, we calculate the *gross mean* ADL per 50-year period, i.e. the mean of all ADLs in a 50-year period, as a coarse-grained measure to detect any major shifts in locality over time (Section 10.1.1). Second, we normalize ADL per sentence length (Section 10.1.2[1]). This means that we observe the ADL per each sentence length (SL) per 50-year period, e.g. we calculate the ADL for all sentences of length 30, etc. In this way, we are able to separate the influence of SL from ADL.

Next, we conduct a micro-analysis inspecting specific SLs, which in the macro-analysis have been shown to be especially frequent in the corpora (Section 10.2). In doing so, we analyze in which of the corpora and at which SL ADL is actually minimized (Section 10.2.1). We furthermore investigate which dependency relations are responsible for the observed changes in ADL (Section 10.2.2). For this, we pick

[1]Part of the research in this section has been published in (Juzek et al., 2020).

the SL with the strongest changes in ADL in the two scientific corpora and analyze the different dependency relations in terms of their frequencies and their DL.

In a final analysis, we aim to address our second hypothesis H1.4b (Section 3.1.2.2) that scientific writing develops towards shorter DL within the construction of RCs, i.e. between the head noun and the embedded verb of the RC. To do so, we calculate the ADL of RCs per SL in order to find out whether RCs on average become shorter over time (Section 10.3.1). Second, we investigate the ADL of specific RC types over time to determine the source of temporal fluctuations of ADL (Section 10.3.2).

We close the chapter with a summary and discussion of our results (Section 10.4).

10.1 Average dependency length per 50-year period

We start by exploring the general development of average dependency length (ADL) calculated in bins of 50-year periods. This coarse measure will serve as an indicator of syntactic complexity and show us whether scientific writing shows a register-specific reduction in syntactic complexity as compared to general language.

10.1.1 Gross average dependency length per 50-year period

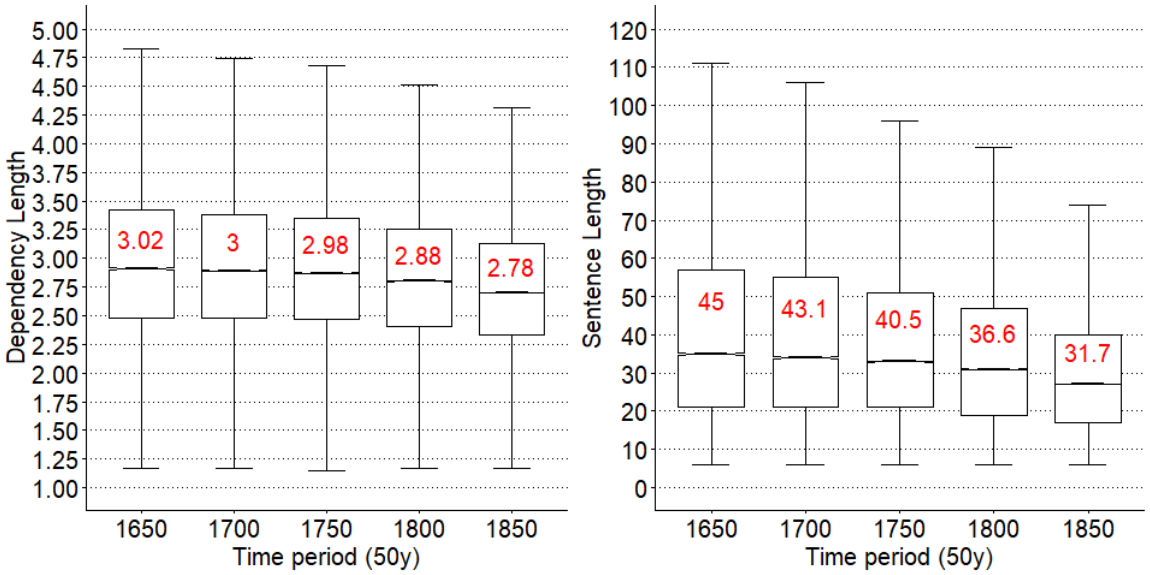
In the present section, we first inspect the development of ADL by observing the coarse overview measure *gross average* ADL, i.e. the mean of all ADLs in a 50-year period. While the full distribution of ADLs in a 50-year period is displayed in box plots, the *gross average* ADL is displayed as numbers in red in the box plots. Additionally, we consider the distributions of the different SLs per 50-year periods to compare the trends in ADL.

10.1.1.1 English

We start by inspecting the distributions of ADL and SL per 50-year period in scientific English (Figure 10.1) and general English (Figure 10.2)[2].

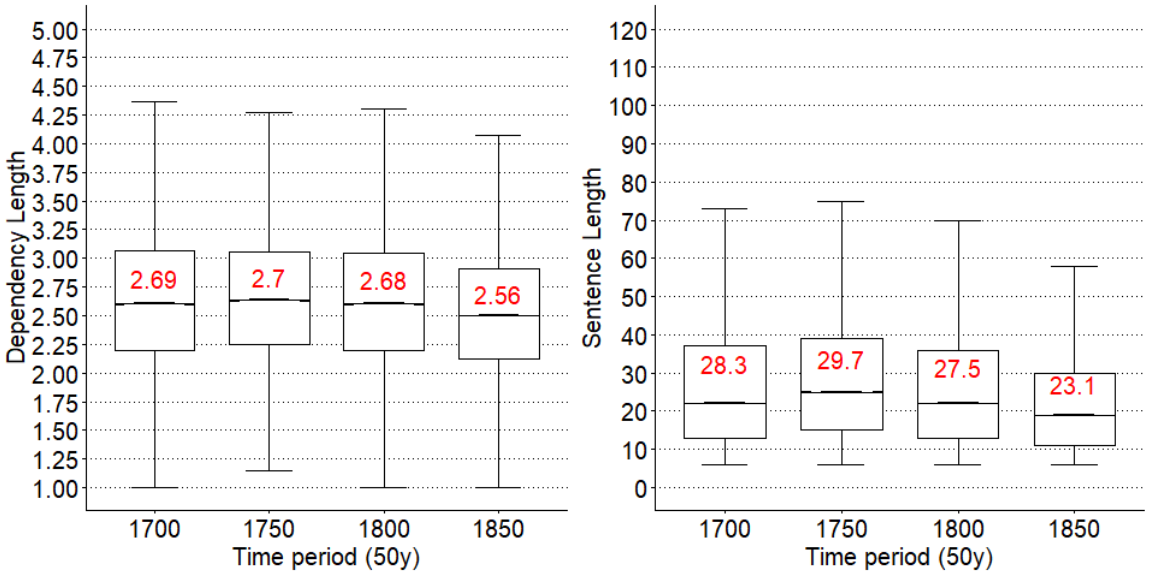
We find that the *gross average* ADL in scientific English (Figure 10.1a) shows a very slight decrease in the first three periods (1650–1750) and a more notable decrease in the last two periods (1800–1850). Apart from ADL, the mean SL also decreases over time (Figure 10.1b). Looking at the first three periods, we can see that the first quartile largely coincides in the SLs 21–34 and only the third quartile covering the longer sentences shrinks, indicating that extremely long sentences become less frequent. In the last two periods, the mean and median SL drop remarkably. Taken together, both SL and ADL show the same trend, suggesting that an overall increasing distribution of shorter sentences seems to be responsible for the parallel downward trend in ADL.

[2]We excluded outliers from the SL box plots.



(a) ADL **(b)** SL

Figure 10.1: Development of (a) ADL and (b) SL in scientific English (RSC) by 50-year periods.



(a) ADL **(b)** SL

Figure 10.2: Development of (a) ADL and (b) SL in general English (CLMET) by 50-year periods.

General English shows a slightly different trend. While there is also an overall decrease in gross average ADL (Figure 10.2a), the decrease is time-shifted: gross average ADL first increases slightly towards the period of 1750 and declines after that, most strongly in the last period, 50 years after the ADL decline we found for scientific English. Also, the mean SL first increases in the 18th and then decreases

in the 19th c. (Figure 10.2b). As in scientific English, the ADL trend runs parallel to SL, suggesting that the gross ADL development seems to be driven by the increasing proportion of short SLs.

Comparing the resulting gross ADL and SL per period, we see that in general English, the decrease is less pronounced than in scientific English. For instance: gross ADL in scientific English decreases by .24 tokens over time, while in general English the decrease amounts to only .13 tokens. The same is true for SL: scientific English sentences decrease by 13.3 tokens while general English sentences decrease by only 5.2 tokens.

Period	ADL	p
1650 vs. 1700	3.02 vs. 2.99	$p < 0.001$
1700 vs. 1750	2.99 vs. 2.98	$p < 0.001$
1750 vs. 1800	2.98 vs. 2.88	$p < 0.001$
1800 vs. 1850	2.88 vs. 2.78	$p < 0.001$

Table 10.1: Gross average ADL per 50-year period and p-values of differences between adjacent periods in scientific English (RSC).

Period	ADL	p
1700 vs. 1750	2.69 vs. 2.70	$p = 1$
1750 vs. 1800	2.70 vs. 2.68	$p < 0.001$
1800 vs. 1850	2.68 vs. 2.56	$p < 0.001$

Table 10.2: Gross average ADL per 50-year period and p-values of differences between adjacent periods in general English (CLMET).

To back up these findings, one-sided t-tests between two adjacent periods (e.g. 1650 vs. 1700) were computed assuming an earlier period has a higher gross average ADL than a subsequent period (see Tables 10.1 and 10.2). We see that in scientific English every gross ADL of a later period is significantly lower than the gross ADLs of their preceding periods. The same holds true for general English with the exception of 1700 vs. 1750, since here gross ADL increases.

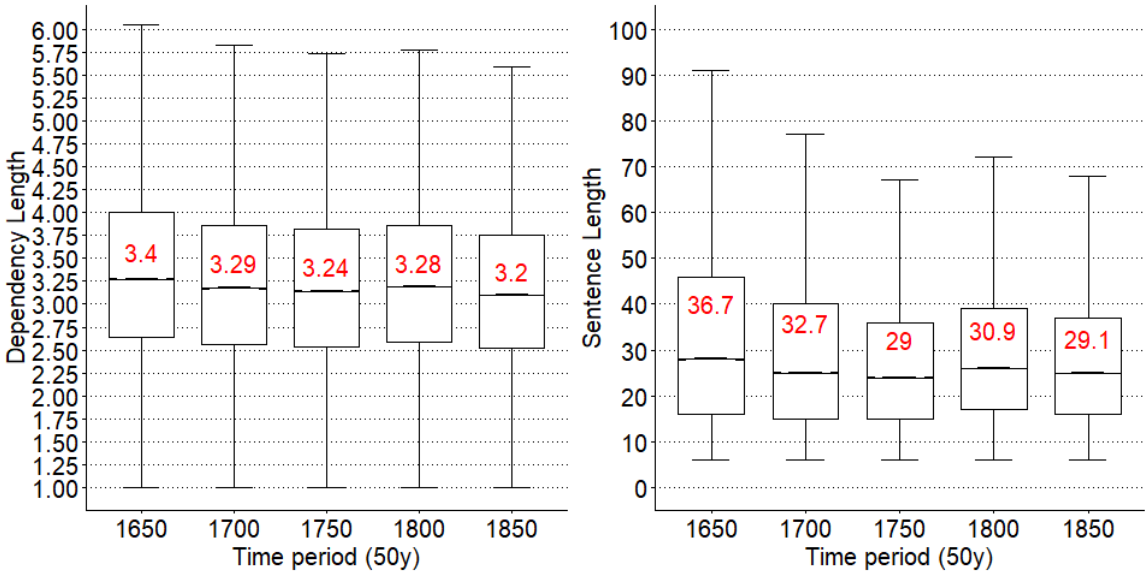
Comparing the decline in gross ADL in both corpora, we see that ADL in scientific English declines more notably from 3.02 to 2.78 than in general English with a maximum ADL of 2.70 declining to 2.56. This shows that the development towards lower syntactic complexity seems to be stronger in scientific English than in general English.

Summarizing the findings up to this point, we have found that in English gross ADL decreases across registers. The decrease in gross ADL seems to be driven by increasingly short sentences, naturally limiting the maximum length of dependency relations and thereby pushing down gross ADL overall. The development, however,

seems to be stronger in scientific English than in general English. In Section 10.1.2, we will test whether ADL still decreases diachronically when normalized per SL.

10.1.1.2 German

Inspecting the gross ADL per period in scientific German (Figure 10.3a) we find a relatively straight downward trend with a bump in 1800. The gross average ADL ranges from a maximum of 3.4 tokens in 1650 and decreases to a minimum of 3.2 tokens in 1850, indicating a clear decline over our observed 50-year periods. Comparing the ADL development to the development of SLs, we see that SLs fluctuate in a similar way as ADLs (Figure 10.3b): the mean SL is at its maximum in 1650 (36.7 tokens) and decreases towards the end of the 18th c., then increases between 1800 and 1849, and decreases again in the last 50 years to a minimum of 29.1 tokens per sentence. Comparing gross ADL and SL in scientific German to the climactic trend in RC frequencies discussed in Chapter 9 (Figure 9.4a), we find divergent trends. While towards the period of 1750–1799, RC frequencies increase steeply, the mean SL in this period is at its minimum and gross average ADL is at its second lowest level. This is astonishing, since one would assume that higher intricacy would also increase ADL. An ad-hoc interpretation of these results could be that the detected decrease in SL is driven by a type of construction other than RCs, such as paratactic structures, falling out of fashion during this time.

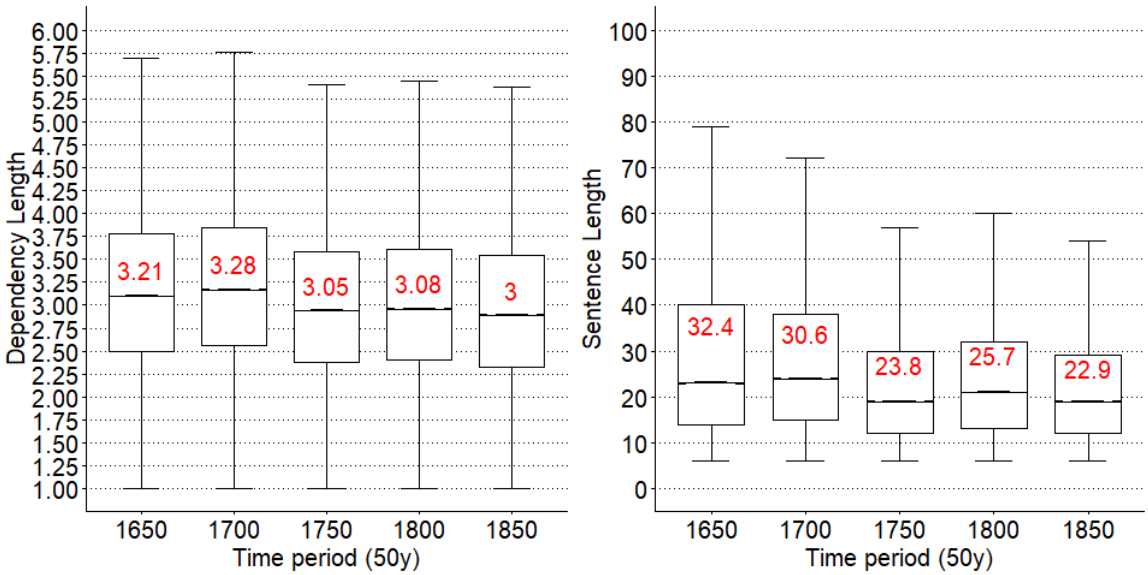


(a) ADL **(b)** SL

Figure 10.3: Development of (a) ADL and (b) SL in scientific German (DTAW) by 50-year periods.

In general German (see Figure 10.4), the gross average ADL per period decreases as well, albeit with stronger oscillations than in scientific German: the gross average

ADL first increases toward its maximum value of 3.28 in 1750 and decreases afterward to reach its minimum of 3 tokens in 1850. The SLs show the same up-and-down trend (Figure 10.4b). The mean SL is highest in 1650 (32.4 tokens) and lowest in 1850 with 22.9 tokens. This decrease of almost 10 tokens on average is remarkable and stronger than that in scientific German, where SL decreases by approximately 7 tokens. Interestingly, the decrease in gross average ADL is almost identical in scientific and general German, with .2 tokens, between 1650 and 1899. The minimum ADL in scientific German is identical to the maximum gross average ADL in general German. At an ADL of 2.21, general German has a SL of 32.28 tokens, while scientific German has a SL of 29.1 tokens. This suggests that scientific German shows higher ADL than general German when holding the SL stable, which is quite possibly due to the higher intricacy of sentences we found in Section 9.1.2. Also, the decrease in gross average ADL is fairly similar between scientific and general German; however, general German seems to show a more intense decrease in SL than scientific German. The general German mean SL decreases by 9.6 tokens, while the scientific German mean SL decreases by 7.6 tokens.



(a) ADL **(b)** SL

Figure 10.4: Development of (a) ADL and (b) SL in general German (DTAG) by 50-year periods.

To determine the significance of the decreases in gross average ADL from one period to the next, we calculate one-sided t-tests comparing all ADLs from one period to the next. The results (see Tables 10.3 and 10.4) confirm that the decreases are significant (while an increase in ADL is non-significant since the one-sided t-test only looks for significant decreases).

Period	ADL	p
1650 vs. 1700	3.40 vs. 3.28	$p < 0.001$
1700 vs. 1750	3.28 vs. 3.24	$p < 0.001$
1750 vs. 1800	3.24 vs. 3.28	1
1800 vs. 1850	3.28 vs. 3.20	$p < 0.001$

Table 10.3: Gross average ADL per 50-year period and p-values of differences between adjacent periods in scientific German (DTAW).

The results show that in scientific German, the gross average ADL decreases significantly in three out of four comparisons. Only between 1750 and 1800 did the gross average ADL increase. General German, meanwhile, shows greater fluctuation.

The results thus far have shown that the development of gross average ADL in German is not as linear as that in English, which is in line with our hypothesis H2. Especially the drop in gross average ADL in the last two periods reflects the time-shifted trend in syntactic complexity reduction we expected to find in scientific German.

Period	ADL	p
1650 vs. 1700	3.21 vs. 3.28	1
1700 vs. 1750	3.28 vs. 3.04	$p < 0.001$
1750 vs. 1800	3.04 vs. 3.07	1
1800 vs. 1850	3.07 vs. 2.99	$p < 0.001$

Table 10.4: Gross average ADL per 50-year period and p-values of differences between adjacent periods in general German (DTAG).

10.1.2 Average dependency length normalized per sentence length

To normalize ADL per SL, the ADL for each sentence length is calculated, i.e. we bin all ADLs of all sentences of the same length and take the mean per 50-year period. The minimum sentence length we consider is eight tokens. Sentences with less than eight tokens are excluded from analysis since shorter sentences tend to include too many incomplete sentences resulting from inaccurate sentence splitting (even after preprocessing the data). The upper bound was set to include only SLs with $n \geq 100$

instances leading to an upper bound of SL = 60. We then plot the ADL per SL. Each colored line represents a 50-year period (e.g., red for 1850–1899). s

10.1.2.1 English

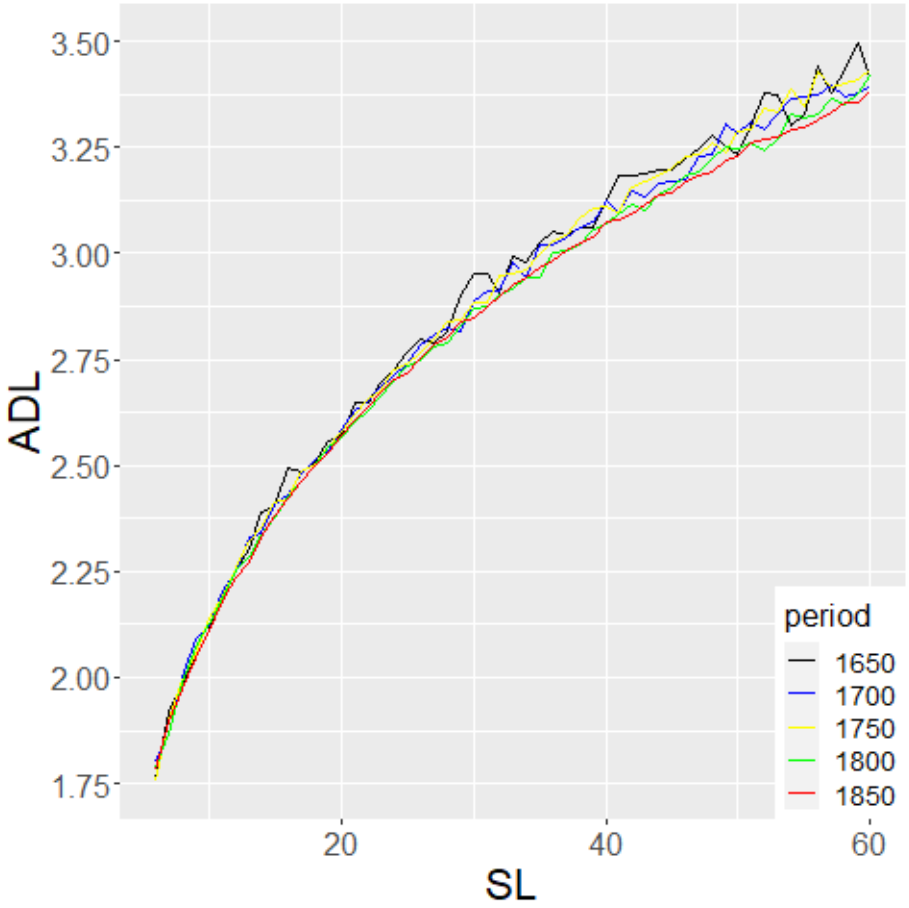


Figure 10.5: Development of ADL per SL in scientific English (RSC) by 50-year periods.

In scientific English (Figure 10.5), the uppermost line indicating the period with the overall highest ADL is the black line representing the earliest period (1650–1699). The lowest line in scientific English indicating the shortest ADL is the red line representing the last time period (1850–1899). The decreasing trend of normalized ADL shows that in scientific English, ADL also decreases when correcting for the bias created by overall trends in SL.

In general English (Figure 10.6), the uppermost (blue) line also represents the earliest period (1700–1749); however, the lowest line is the yellow line representing 1750–1799. Thus, in general English, normalized ADL per SL does not seem to decrease chronologically. The results for gross ADL and normalized ADL in combination show that in general English, ADL does not seem to decrease independently of SL, i.e. equally long sentences over time do not seem to make use of shorter dependencies.

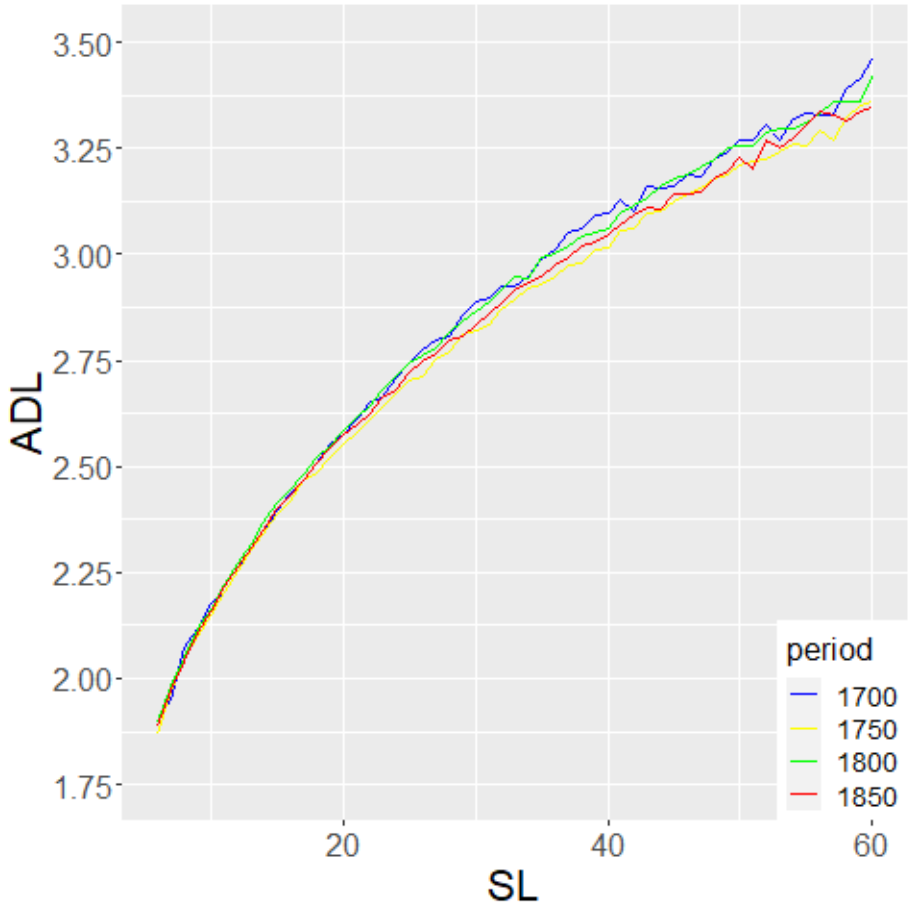


Figure 10.6: Development of ADL per SL in general English (CLMET) by 50-year periods.

What does seem to be happening is that in general English, sentences themselves become shorter, automatically limiting gross ADL over time. In scientific English, meanwhile, the chronological decrease in normalized ADL shows that DL between syntactically related items is not limited by SL but rather by virtue of syntactic processes resulting in shorter DLs per sentence. In addition to this, scientific English SL decreases, too. This shows that scientific English undergoes a trend towards syntactic complexity reduction in terms of two factors: mere SL as well as ADL itself, confirming our H1.4a.

Connecting our findings with the results from our intricacy analysis (Chapter 9), the higher values as well as the more extreme decrease in both gross average ADL and mean SL in scientific English can be attributed to the fact that in scientific English, there is an overall higher number of sentences including RCs as long-distance dependencies compared to general English, as well as a much stronger and more linear decrease of these over time. Also, the downward trend in normalized ADL per SL in scientific English is in line with the decreasing syntactic intricacy we found for scientific English: fewer intricate structures building long-distance dependencies

contribute to a decreasing ADL at the sentence level.

While the graphs in Figures 10.5 and 10.6 give a good overview of the ADL trends in the English corpora, this type of visualization only shows a limited picture. The lines representing the different periods often overlap and fluctuate strongly due to differing sample sizes per SL. This is why it is hard to tell whether all ADL values at every SLs are in line with the overall trend, or if the differences are significant from one 50-year period to the other. In Section 10.2 we will therefore inspect ADL per 50-year period for three highly frequent SLs per corpus.

10.1.2.2 German

Turning to analyze the German development of ADL normalized by SL over time, we find differing chronological trends between the two German corpora.

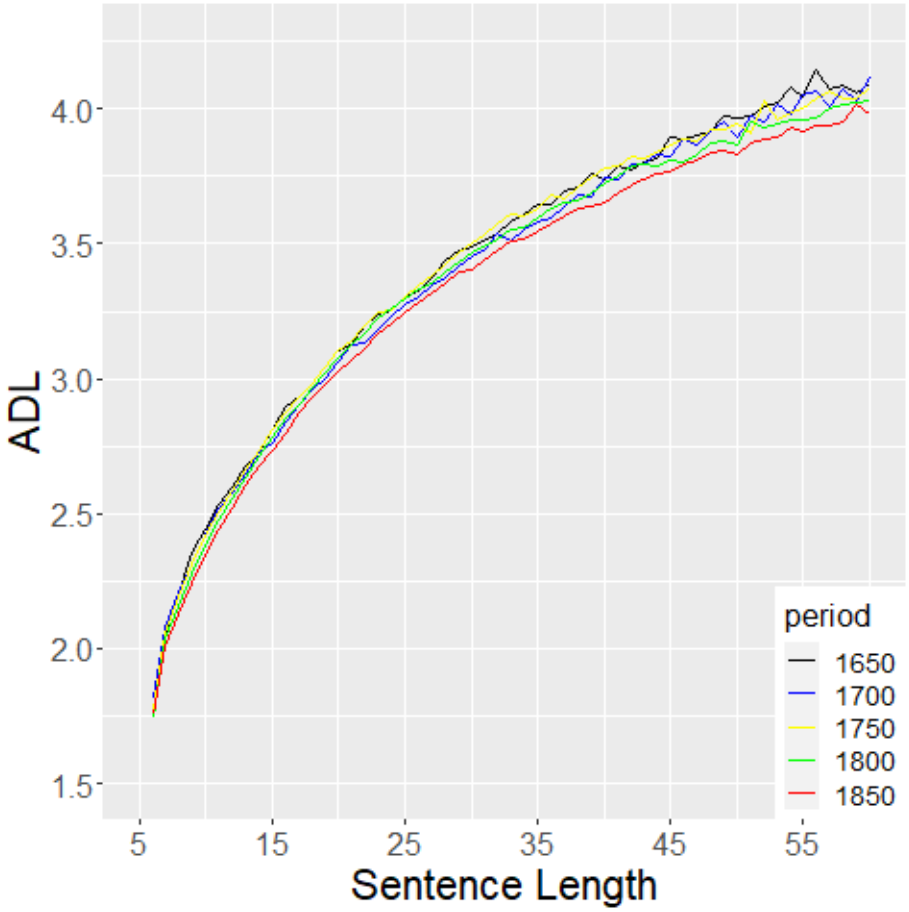


Figure 10.7: Development of ADL per SL in scientific German (DTAW) by 50-year periods.

In scientific German (Figure 10.7) we find the expected decrease in ADL over the observed 50-year periods, i.e. the black line representing the period of 1650 is almost consistently the uppermost line, while the red line representing the latest period is

(with a notable distance from the other periods) the lowest line in the graph. This indicates that ADL independently of SL decreases diachronically in scientific German. In the general German corpus (Figure 10.8), we find a more inconsistent picture: the longer SLs (SL $>$ 30) seem to increase in normalized ADL over time (red is the uppermost line and black the lowest), while the shorter sentences show by far the longest ADL in 1700.

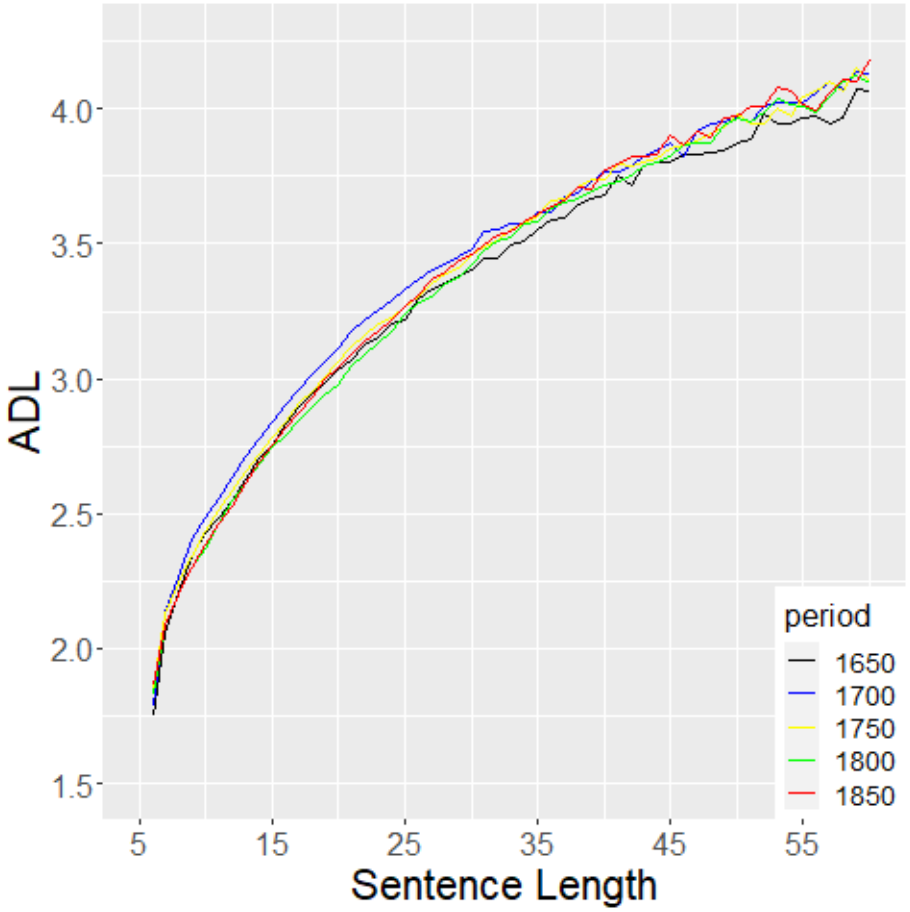


Figure 10.8: Development of ADL per SL in general German (DTAG) by 50-year periods.

The other periods seem to differ little in ADL at shorter SLs, i.e. SL $<$ 30. Surprisingly, the longer SLs show an almost inverse order to that found for scientific German: Here, the black line representing 1650–1699 is the period with the lowest ADL while the red line, which represents the last period (1850–1899), is the uppermost line, i.e. the period with longest ADL.

The developments are similar to those in English in that the scientific corpus seems to decrease chronologically in normalized ADL, while general German does not. Also, the fact that gross ADL decreases in both corpora but is not reflected in normalized ADL per SL in general German suggests that in general German, the decrease in gross average ADL is merely an artifact of decreasing SL. In scientific German, both

shorter sentences and shorter dependency relations together seem to contribute to lower syntactic complexity, which confirms our hypothesis H1.4a.

In summary, the first part of the macro-analysis of German (Section 10.1.1.2) has shown that both German corpora show a decreasing trend in the gross ADL and SL over time. However, much like in English, both ADL and SL are longer in scientific German than in general German. The result is plausible bearing in mind the higher syntactic intricacy in scientific German as identified in Section 9.1.2. At the same time, the decreasing trend in gross average ADL in both corpora seems to be driven by decreasing SL. The second part of the macro-analysis (Section 10.1.2.2) also showed a similar result as that found for English: while in scientific German normalized ADL per SL decreased chronologically, in general German, we did not find such an effect. From both analyses, we can conclude that scientific German seems to increase locality both as a forced effect of decreasing SL and in terms of shorter ADL per SL. As in English, the general German overall drop in gross average ADL per period derives from an overall increase in shorter SLs entailing shorter ADLs.

10.2 Controlling for sentence length

In Section 10.1.1, we have analyzed the overall trend of gross average ADL considering all SLs in the corpora, which is biased by the dominant SLs of each period. To avoid this bias, we will now focus on meaningful decreases in ADL at specific SLs. In Section 10.2.1 we first determine the SLs with significant trends of DLM. In Section 10.2.2, we inspect one specific SL representative of our scientific corpora.

10.2.1 Determining representative sentence lengths

We select the three most common SLs (20, 30, 40) in the scientific corpora (RSC and DTAW) and the three most common SLs (15, 20, 25) in the general language corpora (CLMET and DTAG) to examine their individual developments. We conduct one-sided t-tests to determine if the decrease in average ADL per SL from one period to the next is significant. The p-values are represented in heat maps, with significant p-values ($p < 0.05$) shown in blue and non-significant p-values in gray. The line plots with error bars display the average ADL at each SL and provide information on the significance of the decrease from one period to the next[3].

10.2.1.1 English

For SL20 in scientific English, we do not find any significant differences in ADL over time (Figure 10.9). At SL30, however, all periods have a significantly lower ADL compared to 1650, and 1850 has a lower ADL than all previous periods (Figure 10.10).

[3]Asterisks indicate the significance level: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ and $^{****}p < 0.0001$.

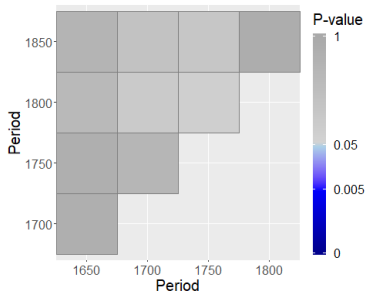


Figure 10.9: RSC SL20

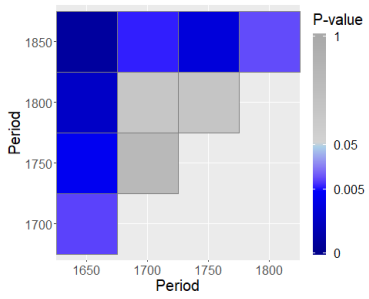


Figure 10.10: RSC SL30

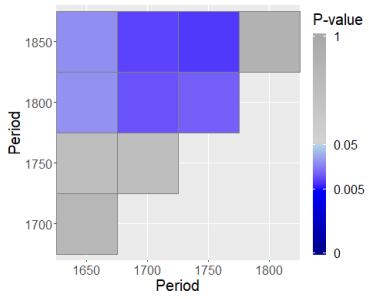


Figure 10.11: RSC SL40

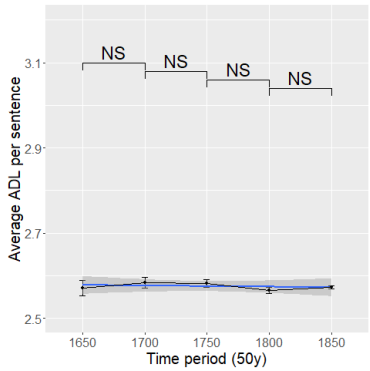


Figure 10.12: RSC SL20

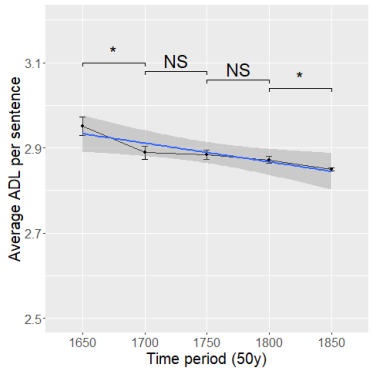


Figure 10.13: RSC SL30

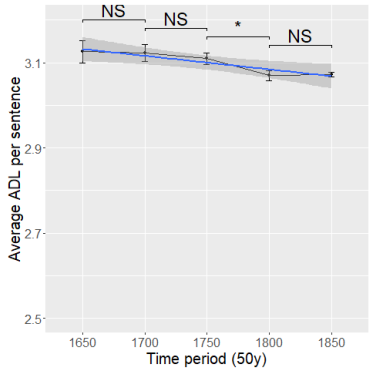


Figure 10.14: RSC SL40

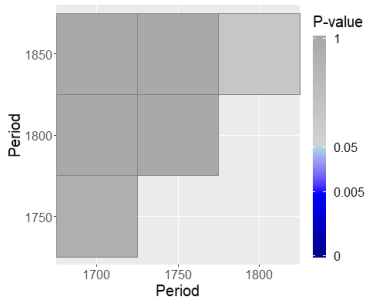


Figure 10.15: CLMET SL15

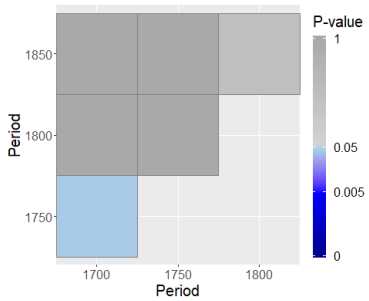


Figure 10.16: CLMET SL20

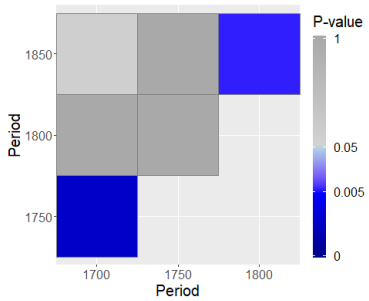


Figure 10.17: CLMET SL25

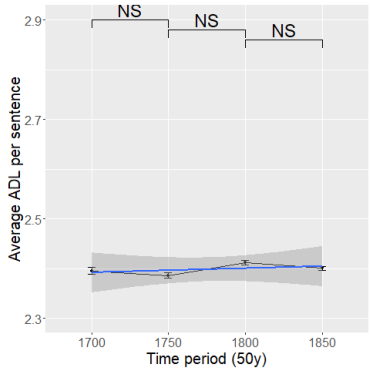


Figure 10.18: CLMET SL15

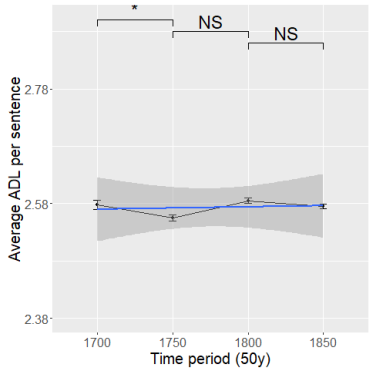


Figure 10.19: CLMET SL20

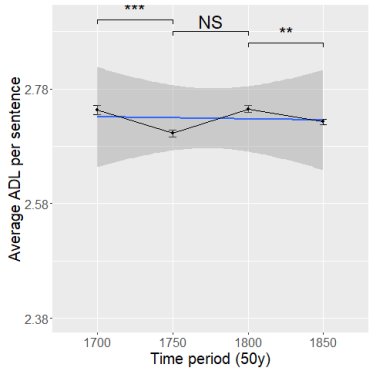


Figure 10.20: CLMET SL25

This development can also be observed by inspecting the corresponding line plot (Figure 10.13), showing a clear downward trend in average ADL at SL30. At SL40 (Figure 10.11), there is a significant drop between the first three periods (1650–1799) and the last two periods (1800–1899). While the trend at SL40 is overall downward, the decrease in ADL seems to slow down in the last 50 years. The heat maps support the findings that the decrease in ADL between 1650 and 1850 is most evident at SL30, where subsequent periods are significantly different from each other between 1650 and 1700 and between 1800 and 1850.

The heat maps for general English (Figures 10.15–10.17) indicate that the ADL remains relatively unchanged over time at all three SLs. The line plots (ADL) (Figures 10.18–10.20) also reflect this stability over the period of observation, resulting in a consistently horizontal trend line at all three lengths. This analysis helps to understand the results from the macro-analysis (Section 10.1), which showed a declining trend in gross average ADL on the one hand, but a non-chronological ordering of ADL per SL on the other. By keeping the length of the SL constant, we have shown that ADL at a stable SL actually does not change over time, and confirmed our suspicion that the decrease in ADL in general English is merely due to the decrease in SL.

Apart from the development of ADL over time in the two English corpora, it is interesting to see whether ADL differs between registers at the same SL. Comparing the results for scientific and general English at SL20, we find that ADL in both corpora is strikingly similar (≈ 2.58). Calculation of a two-sided t-test across the periods covered by both corpora (1700–1850) yields a non-significant p-value ($t = 0.65465$, $df = 3.973$, p-value $= 0.5487$) reflecting the striking similarity in ADL in both English corpora (Table 10.5).

We now have a clearer understanding of how ADL has changed over time in both scientific and general English writing: both registers show a downward trend in gross average ADL. However, in the general language corpus, the decrease in ADL is due to a preference for shorter sentences naturally limiting ADL. In contrast, the scientific data show a chronological decrease in ADL per SL, especially at the highly frequent SL30, showing that in scientific writing, the decrease in ADL is not just an artifact of shorter SL but that there must be syntactic shifts at work leading to a decrease in ADL.

	ADL	
Period	RSC	CLMET
1700	2.58	2.58
1750	2.58	2.55
1800	2.57	2.58
1850	2.57	2.57

Table 10.5: ADL at SL20 in scientific (RSC) and general English (CLMET).

	ADL	
Period	DTAW	DTAG
1650	3.10	3.03
1700	3.06	3.11
1750	3.10	3.06
1800	3.08	2.98
1850	3.03	3.03

Table 10.6: ADL at SL20 in scientific (DTAW) and general (DTAG) German.

10.2.1.2 German

We now turn to determine SLs in which ADL decreases significantly over time in German. In the scientific German corpus (DTAW), all SLs show a significant decrease in ADL in 1850 compared to all other time periods (Figures 10.21–10.23), while the first two periods compared to 1750 do not show a significant decrease in ADL at any of the SLs. Between 1750 and 1800, the decrease in ADL was at first moderate, followed by a sharper decrease between 1800 and 1850. These findings, however, do not align with our findings in Section 10.1.1.2, which revealed that the gross average ADL as well as mean SL per period mostly decreased during the *first* three time periods (1650–1799). These results are interesting considering our results for SLs 20, 30, and 40, which display a delayed decrease in ADL in the last two time periods (1800–1899). As suspected earlier, the decrease in ADL in the 19th c. seems to derive solely from the observed SL reduction. We can therefore deduce that the "actual" decrease in ADL happened much later, in the 19th c. In other words, the reduction in syntactic complexity in scientific German initially occurred in terms of SL reduction and later on the level of locality through a reduction in ADL.

Finally, our findings for general German suggest that the decrease in ADL occurs only in the 18th century, as also demonstrated by the line plots. Despite this interim drop, the trend of ADL remains relatively stable, indicating that there is no significant decrease in ADL from 1650 to 1899. This is similar to what we found for general English and shows that the decreasing gross ADL per period, as found in Section 10.1.1.2, was solely due to the decreasing mean SL over time. However, if we keep SL constant, it shows that the ADL in general German does not decrease. When comparing both German corpora at SL20, we observe that the ADL in scientific German declined markedly from 1650 to 1899 (by 0.7), while in general German, the ADL at SL20 remains unchanged (3.03) between 1650 and 1850. This confirms our hypothesis H1.5 that scientific German specifically decreases in syntactic complexity, as measured by ADL and supports our language-specific hypothesis of a time-shifted complexity reduction in scientific German.

The results of the analysis support our hypothesis that ADL decreases over time in both scientific English and German, even when holding SL stable. However, the

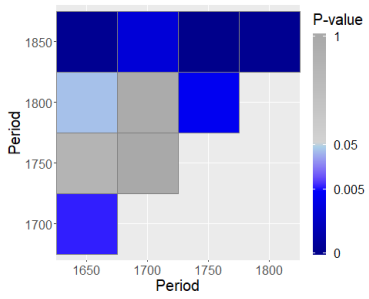


Figure 10.21: DTAW SL20

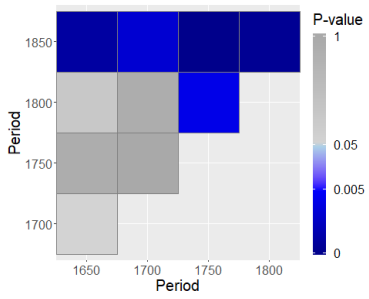


Figure 10.22: DTAW SL30

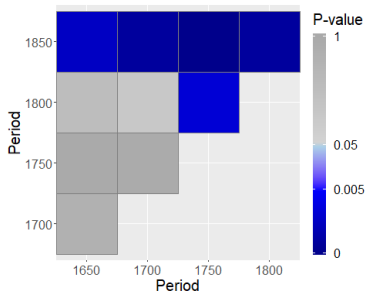


Figure 10.23: DTAW SL40

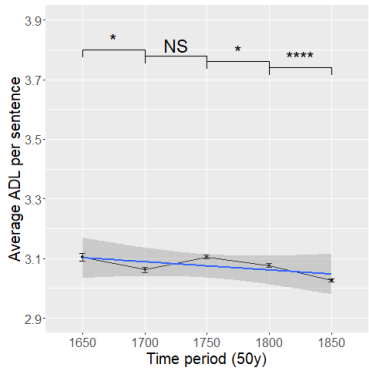


Figure 10.24: DTAW SL20

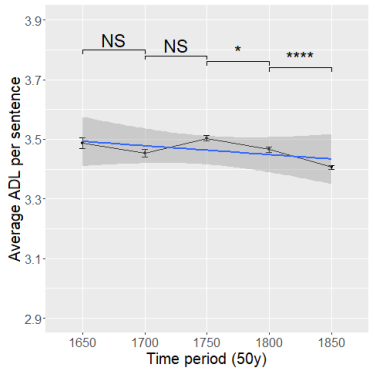


Figure 10.25: DTAW SL30

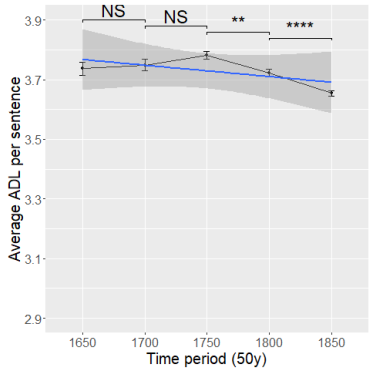


Figure 10.26: DTAW SL40

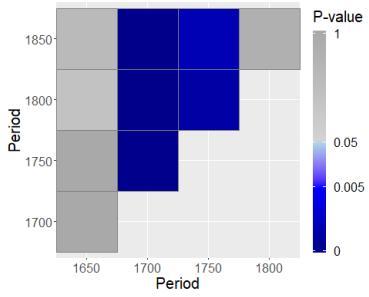


Figure 10.27: DTAG SL15

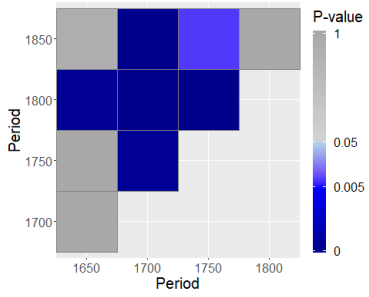


Figure 10.28: DTAG SL20

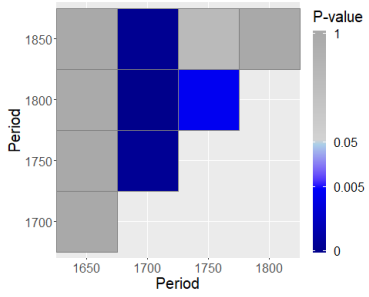


Figure 10.29: DTAG SL25

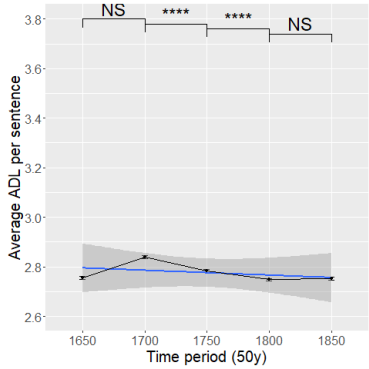


Figure 10.30: DTAG SL15

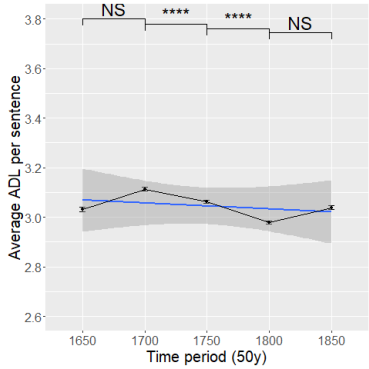


Figure 10.31: DTAG SL20

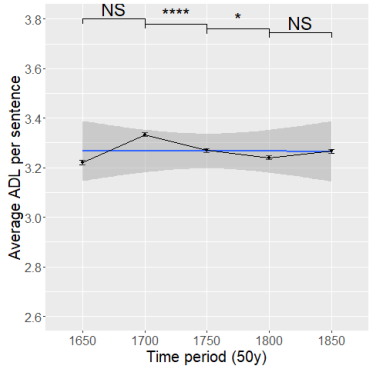


Figure 10.32: DTAG SL25

decrease is more consistent in scientific German, with a reduction in ADL at every analyzed sentence length, while in scientific English the reduction is only highly significant at SL30.

To better understand the changes in ADL, the focus will now be on SL30 in the scientific corpora, where significant decreases were observed in both languages. The aim is to identify factors contributing to the observed reduction of ADL over time.

10.2.2 Analyzing sentence length 30

To examine the factors influencing ADL in the two scientific corpora over time, we focus on comparing the two most distant time periods, 1650 to 1699 and 1850 to 1899, which mark the beginning and end of the development. To gain a deeper understanding of the factors involved in the observed reduction in ADL, we examine the different syntactic relations specified by Universal Dependencies (UD-)relations. More specifically, to determine which syntactic structures contribute most to the overall decrease in ADL, we calculate the ADL[4] and relative frequencies per million tokens (fpm) of each UD-relation type (e.g. `amod` = adjectival modifiers, `nmod` = nominal modifier). To trace the change in ADL and fpm, we furthermore calculate the difference in ADL and the difference in the fpm of each UD-relation between the first and the last period. We combine the two measures ADL and fpm since the overall ADL of a given time period (i.e. the average DL calculated over all UD-relations in a 50-year period) is influenced by two factors:

1. The actual distance between heads and their dependents can increase or decrease; for example, the subject of an object relative clause (RC) can become more complex and lengthen the clause.

2. The proportion of long-distance relations (e.g. inter-clausal relations) and short-distance relations (e.g. nominal phrase components) can change, with long distances becoming less frequent and short distances (such as determiners and adjectival modifiers) becoming more frequent.

High-frequency UD-relations have a greater impact on overall ADL than those with low frequency. To gain a comprehensive understanding of how ADL and fpm changes in particular UD-relations affect the overall ADL trend over time, we calculate four factors that we believe jointly impact the trend:

1. The difference in fpm of each UD-relation in 1850 compared to 1650.

2. The difference in ADL of each UD-relation in 1850 compared to 1650.

3. The overall ADL of a relation binned into three groups: short-distance (< 3), mid-distance (< 6), and long-distance (> 6).

[4] In this analysis the ADL refers to the average DL calculated over all occurrences of a UD-relation, e.g. the summed DLs of all RCs divided by the number of RCs.

4. The average fpm (averaged across all time periods) of a UD-relation, binned into three groups: low-frequency ($< 1,000$ instances per million tokens), mid-frequency ($> 1,000$ instances per million tokens), and high-frequency ($> 10,000$ instances per million tokens).

Using these four measures, we create graphs for the scientific corpora (RSC and DTAW) displaying whether a UD-relation becomes longer or shorter in terms of ADL (x-axis) and whether it becomes more or less frequent (y-axis). The color of each item indicates its average ADL-group (blue for short, green for middle, and red for long), and the size of the label font shows the average fpm group (small for low-frequency, medium for mid-frequency, and large for high-frequency).

10.2.2.1 English

Previous research on scientific English (e.g. Halliday & Martin, 1993; Biber & Gray, 2011b, 2016) suggests that in English scientific texts, noun phrase components (usually head noun adjacent relations) become more frequent, while coordination and clausal post-modification become less frequent over time in the scientific corpus (cf. Section 2.1.3.2). Using the syntactic UD-annotation, these different types of noun phrase modification can be traced (among many other syntactic relations).

In scientific English (Figure 10.33), the strongest increase in frequency is indeed found for short ADL (green), high-frequency (large font) noun phrase internal relations, such as adjectival pre-modifiers (`amod`) with a growth of over 20k compared to 1650, and also for determiners (`det`), case markers (`case`), and nominal modifiers (`nmod`) with an increase of >10k. Moreover, we observe an increase of >5k in oblique nominals (`obl`), passive nominal subjects (`nsubj:pass`), and passive auxiliaries (`aux:pass`), with the latter being the only verbal UD-relation. The frequency of long-distance relations (red) decreases, especially that of conjuncts (`conj`) and parataxis. ADL generally decreases for mid- to high-frequency UD-relations, while it increases for low-frequency UD-relations, with the exception of nominal (passive) subjects (`nsubj/nsubj:pass`), which actually become longer over time. The observations suggest that an overall decrease in ADL over time seems to be due to the gradual preference of UD-relations with rather short ADL and abandonment of long-distance relations. Moreover, highly frequent relations seem to undergo a process of compression as they tend to become shorter over time, whereas low-frequency relations become longer. The development observed here thus points to a restructuring of the syntactic configurations in scientific English based on a trade-off between ADL and frequency leading to an overall shorter ADL.

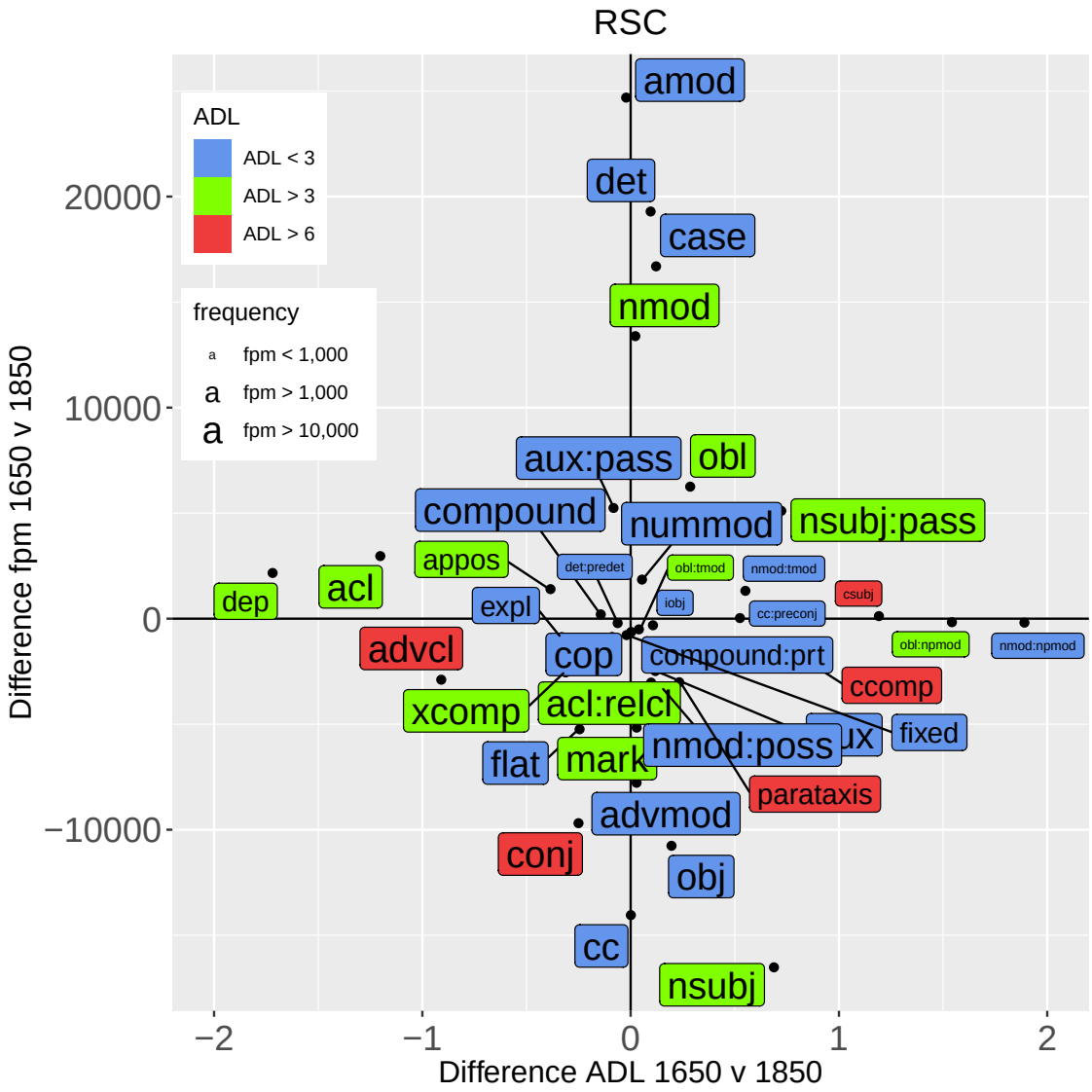


Figure 10.33: Development of ADL (x-axis) and fpm (y-axis) of each UD-relation in scientific English (RSC) in 1650 compared to 1850 at SL30.

Inspecting the development of UD-relations with short, middle, and long ADL in a more aggregate fashion (Figure 10.34), we find that long UD-relations decrease proportionally, while short UD-relations increase significantly ($x - squared = 112.32$, $df = 2$, $p - value < 2.2e^{-16}$). Interestingly, mid-distance UD-relations seem to occupy a stable proportion in scientific English indicating that the increasing preference for short over long UD-relations is the mechanism ultimately driving the trend toward overall shorter ADL in scientific English over time.

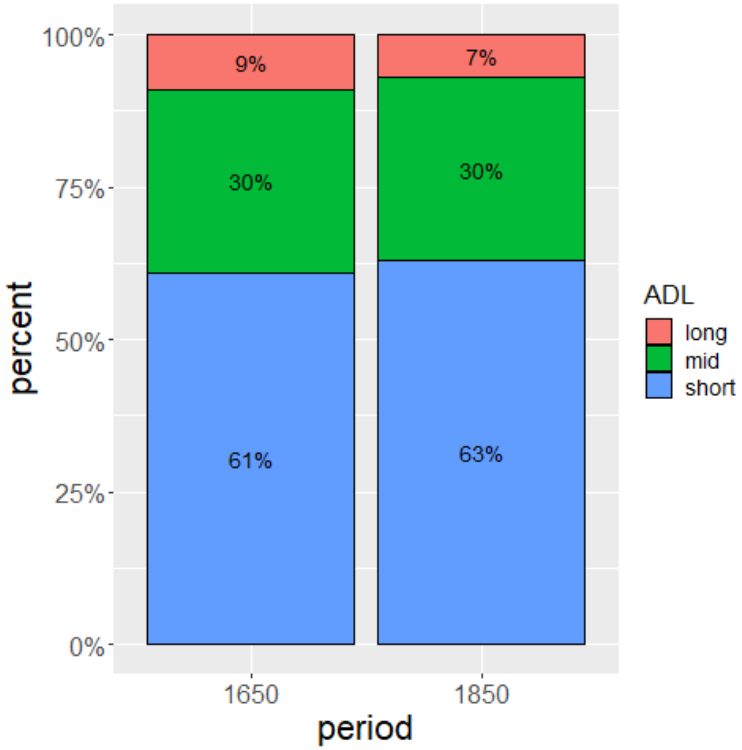


Figure 10.34: Percentage distribution of long (> 6 tokens), mid (> 3 tokens) and short (< 3 tokens) dependency relations in scientific English (RSC).

We finally inspect the percentage difference in fpm of each UD-relation (Table 10.7, column "% Difference"). We observe notable changes in frequency, particularly for some low-frequency UD-relations. Although these low-frequency relations may not greatly impact overall ADL in a period, they could reveal register-specific preferences within scientific writing. In scientific English, short-distance temporal nominal modifiers (`nmod:tmod`) experience a five-fold increase in frequency (see Example (1)), whereas the longer oblique temporal modifiers (`obl:tmod`, see Example (2)) decline by 70%. Additionally, indirect objects (`iobj`) see a decrease of 86%. The increase in temporal modifiers aligns with our expectations since this relation represents a noun or noun phrase that modifies a temporal expression, such as a date or time. It provides additional information about the temporal expression, such as the duration or frequency of an event in an explicit yet compressed form of noun phrase modification convenient in scientific writing. The decrease of the longer oblique temporal modifiers

being a type of adverbial phrase or clause is also in line with our assumptions that in scientific writing shorter and more compressed forms are preferred over long ones. The loss of indirect objects is another interesting observation. A look into the corpus data reveals that indirect objects in the scientific English corpus RSC are overwhelmingly expressed using personal pronouns (Example (3)). Of these personal pronouns in the period of 1650, the pronoun *you* occupies 39.36%, the next most frequent pronoun is *us* with a proportion of 18.08%, and the third most frequent pronoun is *me* with 13.18%. The high usage of first and second-person pronouns is typical of an "involved style" referring to language with an "affective/interactive" focus (cf. Biber, 1988). In the early years, the RSC contained many letters (as in Example (3)) of which an involved style is typical. Over time as the meta-register evolves, letters become less frequent forms of scientific communication and the involved style gradually shifts to an increasingly "informational" style (cf. Biber, 1988). Example (3) not only reflects the use of the indirect object relation but also demonstrates the high degree of intricacy induced by the employment of both parataxis and hypotaxis, resulting in an extremely lengthy sentence.

(1) *The only occasion on which any noteworthy difference would have been produced in the Declinations, was in the case of two observations made with magnetometer 60 on September **30** (head), **1887** (dependent).* (`nmod:tmod`) (A. W. Rucker and T. E. Thorpe, A magnetic survey of the British Isles, 1891)

(2) *In the latter the observation was taken about 9 P.M.; and the next following observation, **taken** (head) between 8 and 9 A.M. the next **morning** (dependent), showed positive electricity of unusual strength.* (`obl:tmod`) (Joseph D. Everett, Account of Observations of Atmospheric Electricity Taken at Windsor, Nova Scotia, 1862)

(3) *Before I go on farther with this History, first I will tell **you**, this Lady had an easy and natural Delivery, and that it was a natural birth, and that the Child came into the World without any force, so that consequently it got not this Wound in its Birth, but was occasioned by strength of Imagination, about two Months before the Mother was gone to Bed, by chance she heard a Report, that a Man had murdered his Wife, and with a Knife had given her a great Wound in her Breast, at which Relation she changed, but not excessively.* (`iobj`) (Part of a Letter from Dr. Cyprianus to Dr. Sylvestre, Giving an Account of a Child Born with a Large Wound in the Breast, 1695)

UD-relation	1650	1850	Difference	% Difference
nmod:tmod	226.43	1547.26	1320.83	583
dep	1358.57	3524.7	2166.13	159
csubj	135.86	258.47	122.61	90
amod	42025.18	66716.24	24691.06	59
nsubj:pass	11547.87	16656.93	5109.06	44
aux:pass	14899.01	20144.46	5245.45	35
acl	11095.01	14061.64	2966.63	27
appos	5298.43	6698.89	1400.45	26
nmod	54025.9	67413.74	13387.84	25
det	103885.52	123178.78	19293.27	19
nummod	10778.01	12631.22	1853.22	17
case	119780.82	136475.65	16694.83	14
obl	51489.9	57744.26	6254.36	12
cc:preconj	543.43	573.58	30.15	6
compound	18748.3	18953.03	204.73	1
cop	17570.87	16704.73	-866.14	-5
ccomp	8332.58	7238.83	-1093.74	-13
advmod	52350.33	44574.86	-7775.47	-15
aux	16257.59	13778.39	-2479.19	-15
mark	33420.89	28259.6	-5161.29	-15
advcl	16846.3	13955.42	-2890.88	-17
xcomp	12589.44	10044.79	-2544.65	-20
conj	46417.9	36728.8	-9689.09	-21
acl:relcl	12544.15	9515.46	-3028.69	-24
compound:prt	2626.57	1996.92	-629.65	-24
det:predet	860.43	651.48	-208.95	-24
nsubj	58056.34	41536.99	-16519.34	-28
obj	37723.03	26961.96	-10761.08	-29
fixed	2581.29	1803.95	-777.33	-30
nmod:poss	10959.15	7702.66	-3256.5	-30
cc	44425.32	30378.67	-14046.65	-32
discourse	90.57	56.65	-33.92	-37
expl	2309.57	1439.27	-870.3	-38
flat	12815.87	7576.96	-5238.9	-41
parataxis	6838.15	3827.43	-3010.72	-44
nmod:npmod	317	129.23	-187.77	-59
obl:npmod	271.71	104.45	-167.27	-62
obl:tmod	724.57	217.75	-506.82	-70
iobj	362.29	49.57	-312.72	-86

Table 10.7: UD-relations and fpm in scientific English (RSC), the difference between 1850 and 1650, and the percent difference between the two time periods for each UD-relation.

10.2.2.2 German

Shifting our focus to the examination of the evolution of scientific German (Figure 10.35), we find a relatively similar development to that in scientific English.

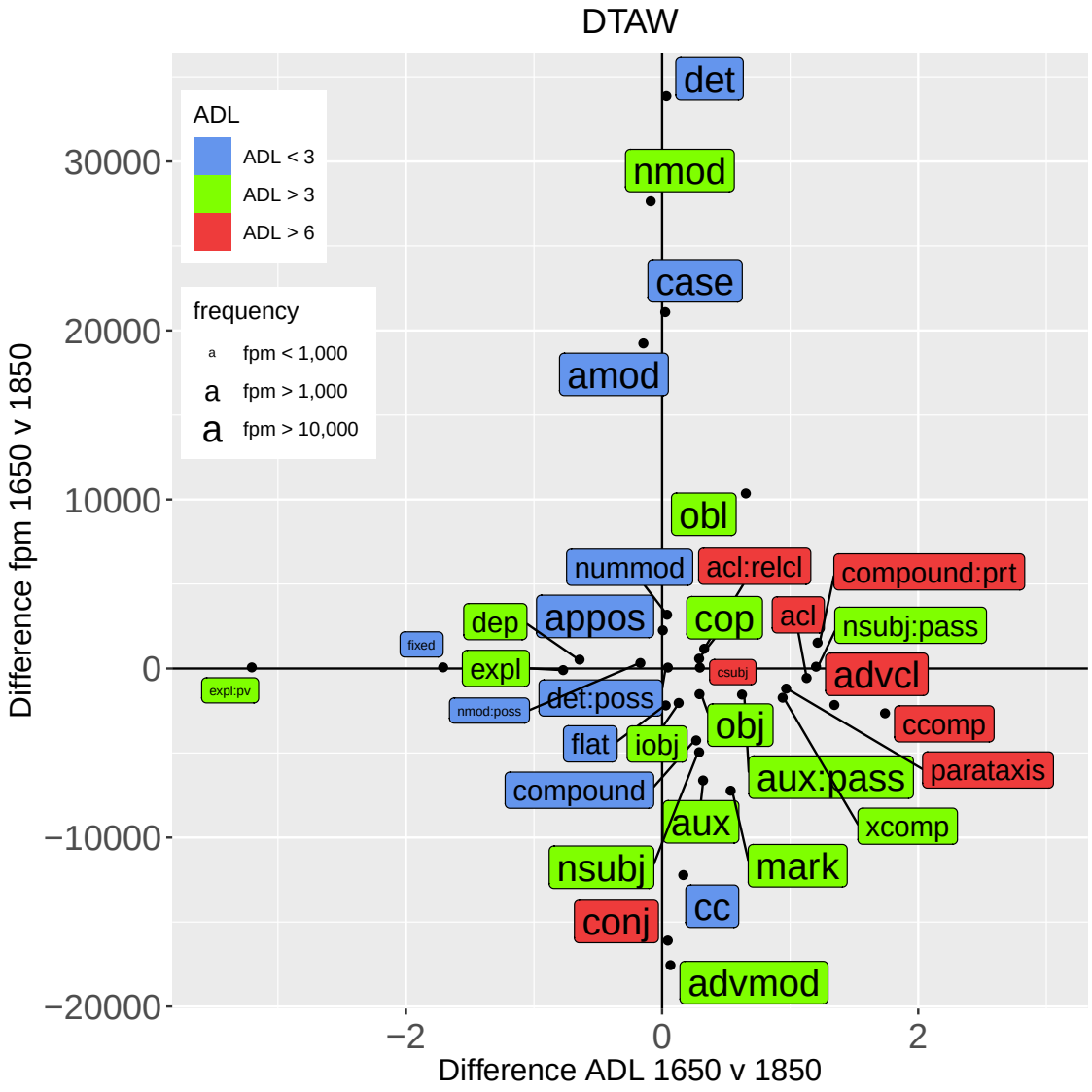


Figure 10.35: Development of ADL (x-axis) and fpm (y-axis) of each dependency relation in scientific German (DTAW) in 1650 compared to 1850 at SL30.

The trend in scientific German also shows a strong increase of the same four high-frequency, short-distance, noun phrase internal relations as in scientific English: `det` ("***die*** *Jochbeine*"), `nmod` ("*Leporiden gleichen* ***Ursprungs***"), `case` ("***in*** *diesen Forschungen*"), and `amod` ("***transversaler*** *Durchmesser*"). The highest increase can be found for determiners (`det`) with an increase of over 30k, followed by nominal modifiers (`nmod`) with an increase of over 25k, case markers (`case`) with an increase of over 20k, adjectival pre-modifiers (`amod`) with an increase of over 15k, and oblique

nominals (obl) with an increase of over 10k.

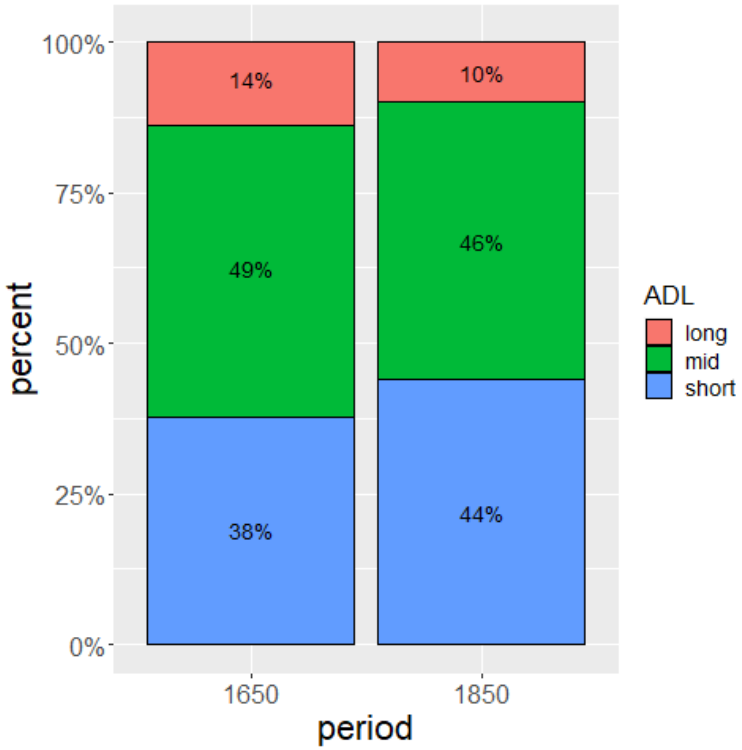


Figure 10.36: Percentage distribution of long (> 6 tokens), mid (> 3 tokens) and short (< 3 tokens) dependency relations in scientific German (DTAW).

Inspecting short-, mid-, and long-distance relations in a more aggregate version (Figure 10.36), the proportional preference for short UD-relations is confirmed. Moreover, unlike in scientific English, both mid- *and* long-distance UD-relations decrease proportionally. The differences between the proportions in the two time periods are highly significant and even more extreme than in English ($x - squared = 1124.4$, $df = 2$, $p - value < 2.2e{-}^{16}$).

While the short-distance UD-relations become significantly more frequent over time, they show little change in terms of ADL. The long UD-relations, however, remain relatively stable in frequency but become longer. In particular, phrasal verb particles (`compound:prt`, Example (4)), adverbial clauses (`advcl`), and clausal complements (`ccomp`) increase in ADL by over 1 token.

In German, phrasal verb particles are obligatorily placed at the end of a sentence, which can lead to long-distance relations, as in Example (4) where an intervening RC increases the distance between the verb *bietet* and its particle *dar*.

(4) *Der menschliche Verstand* ***bietet*** *in der Vollendung, die er der Astronomie zu geben gewusst hat, ein schwaches Abbild solchen Geistes* ***dar****.* (`compound:prt`)
(Emil H. Du Bois-Reymond, Über die Grenzen des Naturerkennens, 1872.)

Adverbial clauses cover constructions such as temporal clauses, consequence, con-

ditional and purpose clauses, etc. In Example (5), the purpose adverbial clause is extremely stretched due to the lengthy noun phrase constructions including a participle construction (*ein noch nicht zinspflichtig gemachtes Volk*) preceding the verb *heranzuziehen* of the adverbial clause.

(5) *Wie es einen Welteroberer der alten Zeit an einem Rasttag inmitten seiner Siegeszüge verlangen konnte, die Grenzen der unübersehbaren seiner Herrschaft unterworfenen Länderstrecken genauer festgestellt zu **sehen**, um hier ein noch nicht zinspflichtig gemachtes Volk zu dem Tribut **heranzuziehen*** [...]. (`advcl`) (Emil H. Du Bois-Reymond, Über die Grenzen des Naturerkennens, 1872.)

Moreover, complement clauses become increasingly long. Example (6) shows a long-distance dependency relation between the head *sieht* and the embedded verb *abnehmen* of the complement clause created by a participle construction premodifying the noun *Arbeitsgrössen*.

(6) *Man **sieht**, dass die nach der Pambour'schen Theorie berechneten Arbeitsgrössen mit wachsendem Volumen schneller **abnehmen**, als die nach unseren Gleichungen berechneten* [...]. (`ccomp`) (Rudolf Clausius: Über die Anwendung der mechanischen Wärmetheorie auf die Dampfmaschine, 1856)

Note that RCs also seem to become slightly longer in terms of ADL. Example (7) illustrates that the RC includes two complex prepositional phrases (*durch gleichartige Verursachung* and *unter den mehrbesprochenen Modifikationen*) richly modifying the head noun *Einzelwerten*. The latter of the two prepositional phrases again includes a participle construction (*den mehrbesprochenen*), suggesting that especially these participle constructions might be responsible for the increase in ADL in some of the low-frequency but long-distance UD-relations found in the scientific German data. The finding that RCs seem to become longer over time is not in line with our hypothesis. We will therefore come back to it separately in the next section (10.3), as well as in Chapter 11.

(7) *Es besteht darin, dass die Gruppierung einer größeren Zahl von **Einzelwerten**, die durch gleichartige Verursachung unter den mehrbesprochenen Modifikationen zu stande **gekommen** sind, zutreffend beschrieben werden kann durch eine mathematische Formel, das sogenannte Fehlergesetz, welche besonders dadurch charakteristisch ist, dass sie nur eine einzige Unbekannte enthält.* (`acl:relcl`) (Hermann Ebbinghaus, Über das Gedächtnis, 1885)

The trends found for scientific German align with our findings for scientific English in that high frequency, short dependency length noun phrase internal relations further increase in frequency. However, we also found that long-distance relations seem to become even longer over time. This seems to be due to the increasing use of richly pre-modified noun phrases (especially through participle constructions) stretching

subordinate clauses. This finding suggests a trade-off between the preference for shorter dependencies on the one hand and an extension in ADL of fewer but longer constructions on the other.

Lastly, we inspect the percentage increases and decreases of the different UD-relations from 1650 compared to 1850 in the scientific German corpus (Table 10.8). We find the greatest increase in percentage in the use of possessive nominal modifiers (`nmod:poss`, +123%). However, looking at the corpus data, we find that, in fact, the label `nmod:poss` in the German parses seems to be consistently erroneously assigned to the genitive, plural form of relativizer *die*, i.e. *deren* (Example (8)). However, since the relative frequency of this relation is very low, we believe that the erroneous annotation does not have a significant impact on our overall results for ADL.

(8) *Es sei t eine Kugel,* ***deren*** *Halbmesser = R ist, und* ***deren*** *Mittelpunkt mit dem Anfangspunkte der Koordinaten zusammenfällt.* (`nmod:poss`) (Carl F. Gauss, Allgemeine Lehrsätze in Beziehung auf die verkehrten Verhältnisse des Quadrats der Entfernung wirkenden Anziehungs- und Abstossungs-Kräfte, 1840)

Numeric modifiers (`nummod`, +89%, Example (9)) add information close to the nominal head and show a notable percentage increase. The percentage increase of this relation is plausible in scientific texts since it conveniently serves to add numerical information in a compressed form.

(9) *Ich finde bei* ***34*** *Hasenschädeln, deren durchschnittliche Basilarlänge = 76,6 mm.* (`nummod`) (Hermann Engelhard von Nathusius, Über die sogenannten Leporiden, 1876)

Moreover, fixed multi-word expressions (`fixed`, +89%, Example (10)) show a strong percentage increase. The label of multi-word expression in UD is relatively loosely defined as "used to connect tokens of a fixed expression"[5]. A look into the corpus data reveals that in the German texts, the UD-relation `fixed` seems to be used mostly for fixed prepositional phrases such as "*von da an*", "*von nun an*", etc.

(10) a. ***Von da an*** *nimmt die Gesamtzahl nicht ab , sie bleibt nur stabil ; die häuslichen Stühle machen 1846 86,1% , 1861 - 86,0% aller auf Leinwand gehenden Stühle aus.* (`fixed`) (Gustav Schmoller, Zur Geschichte der deutschen Kleingewerbe im 19. Jahrhundert, 1870)

b. ***Von nun an****, in dem Tageslichte der nahen Vergangenheit und Gegenwart, zeigt sich uns immer schärfer und deutlicher die Natur dieser und aller Geschichte.* (`fixed`) (Friedrich Theodor von Vischer, Ästhetik oder Wissenschaft des Schönen, 1846.)

[5]Documented at Universal Dependencies.org (2023a).

The strongest percentage decrease in usage is seen in compounds (−75%). This decrease can be explained by multi-word expressions that were previously joined by a hyphen (Example (11)) and over time have become lexicalized into single words.

(11) *Es ist bekannt genug daß alle Vegetabilien, als Korn, Hecken- und* ***Baum-Früchten****, ja alles Gras und Kräuter, durch vorhergehende Präparation und Fermentation, einen Spiritum ardentem geben, aber immer eins mehr und besser als das ander, nach dem es reif oder unreif, fett oder mager in seiner Natur ist.* (`compound`) (Johann Rudolph Glauber, Annotationes, 1650)

To summarize, our findings for scientific German exhibit a similar trend to those found for scientific English, where high-frequency, short dependency length noun phrase internal relations continue to increase in frequency. However, our analysis also revealed that long-distance relations such as subordinate clauses (e.g. RCs) are becoming even longer over time, potentially due to the growing usage of noun phrases with rich pre-modifiers, particularly participle constructions, which stretch subordinate clauses. These results suggest that there may be a trade-off between the preference for shorter dependencies and an increase in the average dependency length of fewer, yet longer constructions.

UD-relation	1650	1850	Difference	% Difference
nmod:poss	263.46	588.21	324.75	123
fixed	76.84	145.14	68.3	89
nummod	3556.65	6737.64	3180.99	89
nmod	37377.74	65022.03	27644.29	74
expl:pv	98.8	165	66.21	67
compound:prt	2316.21	3843.97	1527.76	66
amod	46302.29	65539.96	19237.67	42
det	100069.16	133932.34	33863.18	34
case	72757.61	93847.21	21089.6	29
acl:relcl	4588.52	5758.31	1169.8	25
appos	9550.26	11809.96	2259.71	24
obl	42394.37	52755.25	10360.88	24
csubj	241.5	294.87	53.37	22
dep	2963.87	3489.52	525.64	18
cop	10867.54	11452.46	584.92	5
csubj:pass	21.95	22.92	0.96	4
nsubj:pass	6070.45	6176.93	106.48	2
det:poss	8803.8	8859.77	55.96	1
obj	37882.7	36366.44	-1516.25	-4
acl	10329.65	9764.23	-565.42	-5
nsubj	72318.52	67364.16	-4954.36	-7
expl	1152.62	1058.77	-93.85	-8
aux:pass	10516.26	8974.35	-1541.91	-15
advmod	100453.36	82903.51	-17549.85	-17
advcl	11701.81	9542.7	-2159.12	-18
xcomp	9297.78	7570.29	-1727.49	-19
conj	64063.58	47967.1	-16096.48	-25
flat	8628.17	6444.3	-2183.87	-25
mark	29013.03	21783.5	-7229.53	-25
cc	45138.7	32916.65	-12222.05	-27
iobj	7607.28	5567.34	-2039.94	-27
aux	20308.02	13686.11	-6621.91	-33
ccomp	7091.34	4442.87	-2648.47	-37
parataxis	3227.33	2036.57	-1190.76	-37
compound	5653.31	1401	-4252.31	-75

Table 10.8: UD-relations and their frequency per million tokens in scientific German (DTAW), the difference between 1850 and 1650, and the percent difference between the two time periods for each UD-relation.

10.2.2.3 Summary

The overall decrease in ADL at SL30 in both German and English is primarily caused by a significant rise in short-distance, high-frequency UD-relations, which remain relatively stable in ADL. The sharper decrease in ADL in scientific English (cf. Figure 10.13) can be attributed to the decline in the frequency of all long-distance relationships, including adverbial clauses (`advcl`) and conjuncts, in addition to their decrease in ADL over time. The decrease in ADL in scientific German is less pronounced compared to English as seen in Figure 10.25. The qualitative analysis has shown that the more moderate decrease in German can be attributed to an interplay between the long-dependency relations becoming longer and short-dependency relations becoming more frequent. Also, some of the long-dependency relations even become slightly more frequent, such as phrasal verb particles (`compound:prt`, +1528 instances per million tokens) and RCs (`acl:relcl`, +1170 instances per million tokens). Observing the latter two UD-relations in scientific English, we find quite the opposite: In English, phrasal verb particles and RCs belong to the short and mid-size UD-relations (Figure 10.33), while in German they belong to the long UD-relations (Figure 10.35). The difference between the ADLs of the relations in the two languages is due to word order, which is V-2 in all clause types in English, while in German the finite verb is positioned at the end of subordinate clauses. This affects the dependency length between the head noun and verb in RCs and the distance between the verb and the particle in phrasal verb particles, as English places the verb and particle close together while German positions the particle last in the sentence, thus creating long-distance dependencies.

In both languages, most of the UD-relations cluster in the "+ADL/ −FPM" quadrant, indicating that many dependency relations seem to become longer while becoming rather infrequent over time. Thus, the main impact on the overall ADL reduction seems to come from the frequency dimension where we have found four high-frequency, short-dependency UD-relations becoming increasingly frequent, and high-frequency, long-dependency UD-relations (mostly noun phrase post-modifiers) decreasing in frequency.

The results of our quantitative analyses have shown that ADL in English and German scientific writing has decreased in the period between 1650 and 1900, confirming our hypothesis that scientific writing develops toward stronger locality over time (H1.4a), albeit only at the highly frequent sentence length 30. We have also shown that in general language syntactic complexity is not decreased on the level of ADL but rather only via SL reduction. Our qualitative analysis focusing on SL30 has shown that ADL in both languages is modulated by a clear preference for short-distance nominal modifiers such as attributive adjectives and determiners, while disfavoring long-distance dependencies. We found that ADL decreases to a stronger extent in scientific English than in scientific German. The underlying syntactic developments responsible for the slightly milder decrease in ADL in German can be attributed to

an actual increase in DL of long-distance dependencies such as RCs, phrasal verb particles, and adverbial clauses.

10.3 Dependency length of relative clauses

In the previous sections of this chapter, we have shown that in line with our hypothesis, scientific writing over time develops towards stronger locality due to the increasing reliance on short-distance UD-relations. In contrast to our hypothesis H1.4b, our qualitative analysis however also indicated that RCs belonging to the mid-distance relations (> 3 tokens) in English, and long-distance relations (> 6 tokens) in German, become rather longer in ADL. Since the micro-analysis only looked at SL30 and only in the period of 1650 compared to 1850, we will now investigate the development of the ADL of RCs more closely. For this, we calculate the ADL of all RCs per SLs of the sentences they occur in. We set the upper bound of SL included in the analysis to 150 since SLs of sentences containing RCs are distributed across a large range of SLs.

10.3.1 Average dependency length of relative clauses normalized per sentence length

To calculate ADL of RCs normalized per SL, we proceed as for the whole set of sentences (Section 10.1.2), i.e. we calculate the average DL of all `acl:relcl` relations in relation to the SL of the sentence they occur in. Since we now know that over time SL decreases remarkably, by normalizing for SL we can ensure that we avoid the bias of SL when comparing ADL of RCs across time.

10.3.1.1 English

Scientific English (Figure 10.37) shows an interesting, yet expected result: The order of the colored lines is exactly inverse to that for overall ADL in scientific English (Section 10.1.2, Figure 10.5), indicating that RCs become continuously longer in ADL. This outcome was expected, since in Section 10.2.2 we already discovered that RCs have become longer in the period of 1850 compared to the period of 1650. The graph also shows a non-trivial relationship between SL and ADL; the longer a sentence, the longer the embedded RC, and the longer the embedded RC, the longer the overall sentence. Interestingly, general English (Figure 10.38) does not show the diachronic effect: the lines mostly overlap for SLs up to 60 and show inconsistent trends for the longer SLs probably due to a scarcity of data points. However, the relationship of longer RCs leading to longer sentences, or vice versa, does seem to hold in general English as well.

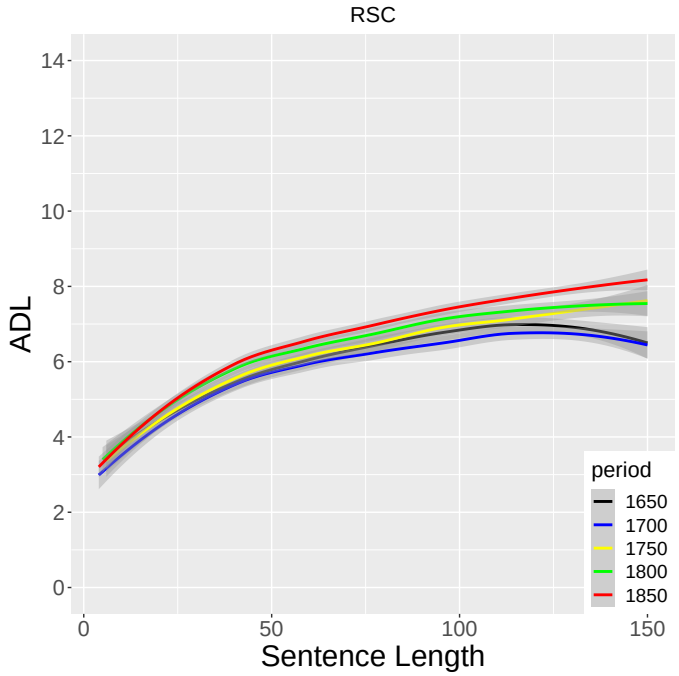


Figure 10.37: ADL of RCs per SL per 50-year period in scientific English (RSC).

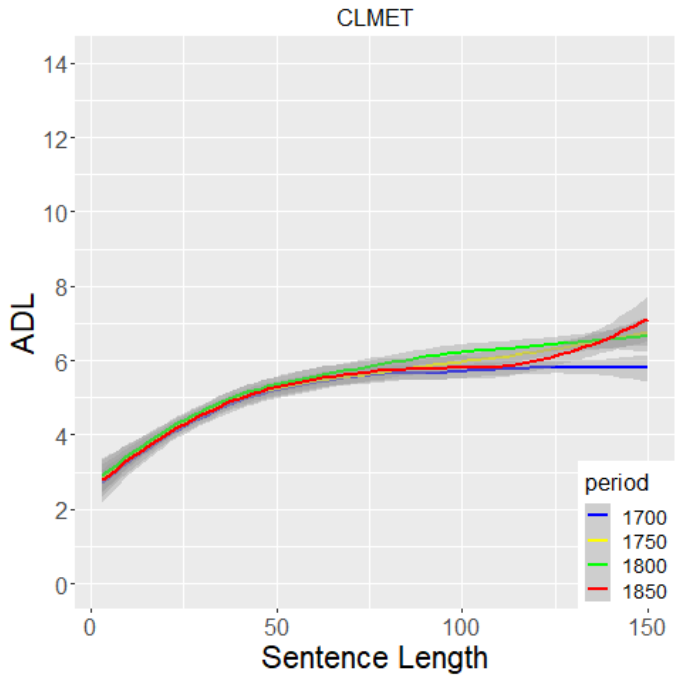


Figure 10.38: ADL of RCs per SL per 50-year period in general English (CLMET).

10.3.1.2 German

For scientific German (Figure 10.39), we find a similar trend as in scientific English. Overall, RCs seem to become longer in ADL, since the red line is the uppermost line at almost all SLs. However, in German, there seems to be a less linear trend toward

longer RCs since RCs initially become shorter from 1650 to the period of 1700 and then continuously longer. Looking at the general German data for comparison (Figure 10.40), we find that RCs almost continuously become longer in ADL, with a slight drop in 1800. Moreover, the trend reverses for longer SLs: in sentences with more than 100 tokens, RCs actually become shorter in the period of 1850. The developments in the German corpora stand in contrast to the overall decreasing development of ADL, but this is in line with our findings from Section 10.2.2, where RCs at SL30 increased from 1650 compared to 1850. Our results from Chapter 7 indicated that in scientific German, RCs defining topicalized head nouns especially increase in frequency. When the head noun is topicalized, the distance between it and the embedded verb of the RC is naturally extremely long (compare Figure 10.41).

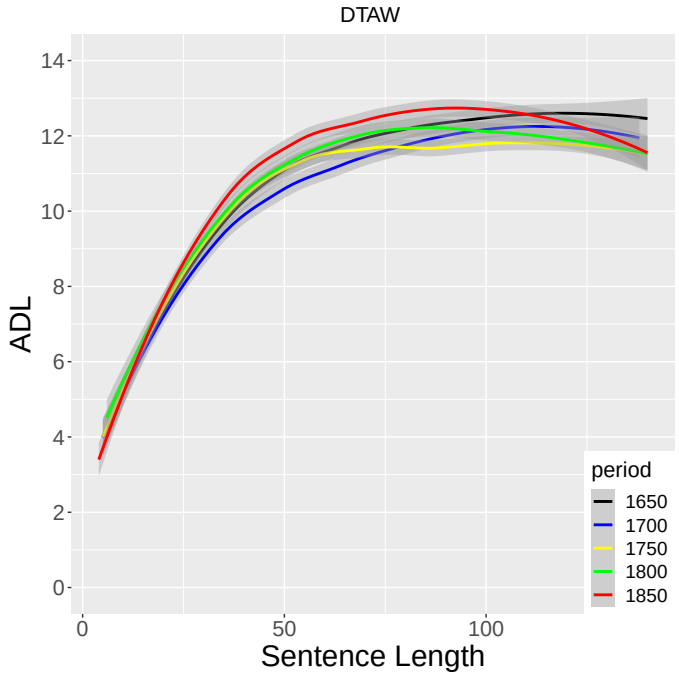


Figure 10.39: ADL of RCs per SL per 50-year period in scientific German (DTAW).

The analysis focusing on RCs separately has thus disconfirmed our hypothesis H1.4b assuming that in scientific writing RCs will show a trend toward shorter DL between the head noun and the embedded verb on average. Instead, RCs seem to create increasingly long dependencies. We now would like to detect the underlying mechanisms that drive the increase in the ADL of RCs. To do so, we will inspect different RC types and their respective ADLs, assuming that, for instance, oblique RCs create longer dependencies since they have to accommodate at least one more token between the nominal head and the embedded verb of the RC (Figure 10.42).

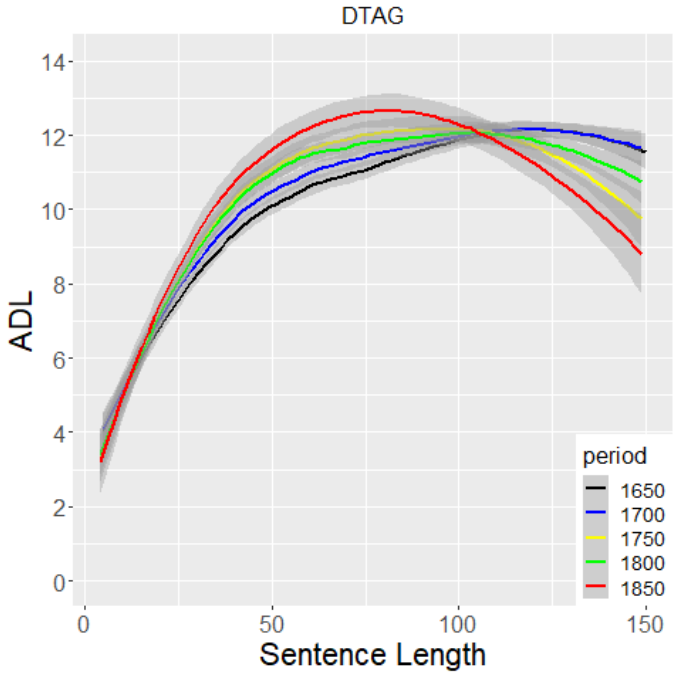


Figure 10.40: ADL of RCs per SL per 50-year period in general German (DTAG).

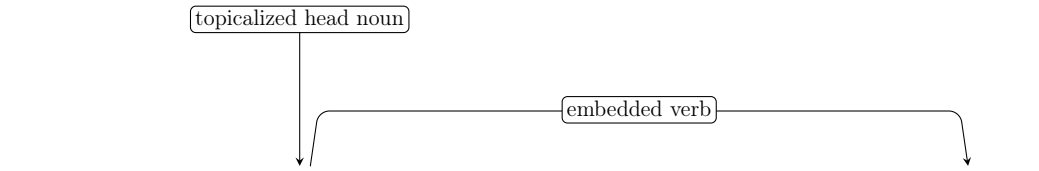


Figure 10.41: Topicalized head noun and intervening material stretching DL (Houston Stewart Chamberlain, Die Grundlagen des Neunzehnten Jahrhunderts, 1899).

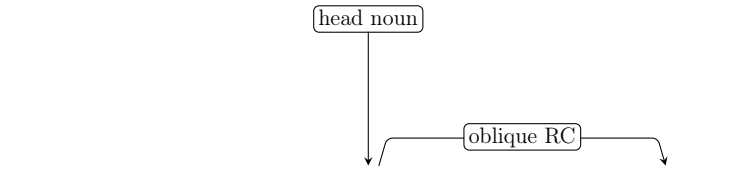


Figure 10.42: Oblique RC and intervening material stretching the DL (Henry Brougham and Charles Blagden, 1796).

10.3.2 Average dependency length per relative clause type

To determine the syntactic extraction type of each RC, we extract the relativizers from the corpora, as they are UD-annotated with their respective syntactic function, i.e. subject (`nsubj`), passive subject (`nsubj:pass`), object (`obj`), indirect object (`iobj`), oblique (`obl`) – note that the English UD-annotation does not distinguish between

indirect objects and oblique nominals. We then calculate the DL (i.e. the distance between the head noun and embedded verb) for each RC and average the obtained values per 50-year period and per RC type yielding the ADL of each RC type per 50-year period. The resulting graphs show the development of ADL of each different RC type per 50-year period.

10.3.2.1 English

For scientific English (Figure 10.43), we find that of all RC types, oblique RCs create the longest dependency relations by far. While all RC types seem to decrease in ADL over time, oblique RCs (purple line) show the strongest decrease, but remain as the longest RC type of all with on average at least 2 tokens higher ADL than the other types. All other RC types are below 6 tokens in length on average, with subject RCs being the overall shortest type. The fact that all RC types seem to become overall shorter over time is astonishing, since our analysis of ADL of RCs normalized by SL (Section 10.3.1) indicated that RCs become longer and not shorter. So how is it possible that ADL of RCs increases over time but individual types of RCs decrease in ADL? Again here, we may assume that the overall ADL as an aggregate measure not only reflects a trend in whether RCs become longer or shorter, but also whether longer and shorter types become more or less frequent. In Chapter 7, we found a strong increase in the pattern DETERMINER NOUN PREPOSITION preceding RCs in scientific English. The lexico-grammatical pattern represents the oblique RC type (Example 10.42), which creates longer distances than other RC types due to the obligatory preposition preceding the relativizer. We may thus assume that an increase in oblique RCs has an influence on the overall ADL increase detected for RCs. Since this question is central to the chapter on accessibility, we will analyze the frequencies of the different RC types in the next Chapter (11).

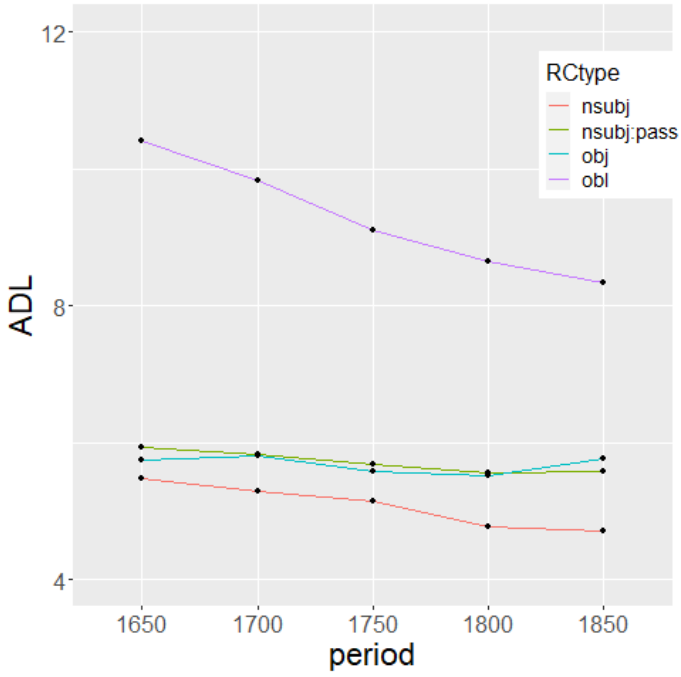


Figure 10.43: ADL per RC types per 50-year periods in scientific English (RSC).

10.3.2.2 German

In scientific German (Figure 10.44), the differences in ADL between the different RC types are less pronounced than in English, which can be explained by the German verb-last word order in subordinate clauses. However, it is surprising that subject RCs do not seem to be the shortest variant, even though in subject RCs theoretically the verb can directly follow the relativizer (as indicated in Example (12) by the optional direct object *den Hund* in brackets) making subject RCs the theoretically shortest RC type.

(12) a. Die Frau **die** (den Hund) schlug [...]. (subject RC)
b. Die Frau **die** der Hund biss [...]. (direct object RC)
c. Die Frau **der** der Hund vertraut [...]. (indirect object RC)
d. Die Frau mit **der** der Hund spielt [...]. (oblique RC)

In all other RC types, at least a subject intervening between the relativizer and the verb is necessary. In our data, however, direct object RCs show the shortest ADL followed by indirect object RCs.

The overall trend of ADL of the different RC types in scientific German is decreasing as in English. However, the ADL of all RC types reaches a minimum in 1750 and increases slightly afterward. As in English, the downward trend in ADL does not explain the overall increase in RC ADL; instead, we seem to encounter an interplay of overall declining DL of RCs and possibly a shift in frequencies of the differently long types, presumably an increase in more extended types (i.e. topicalized head nouns)

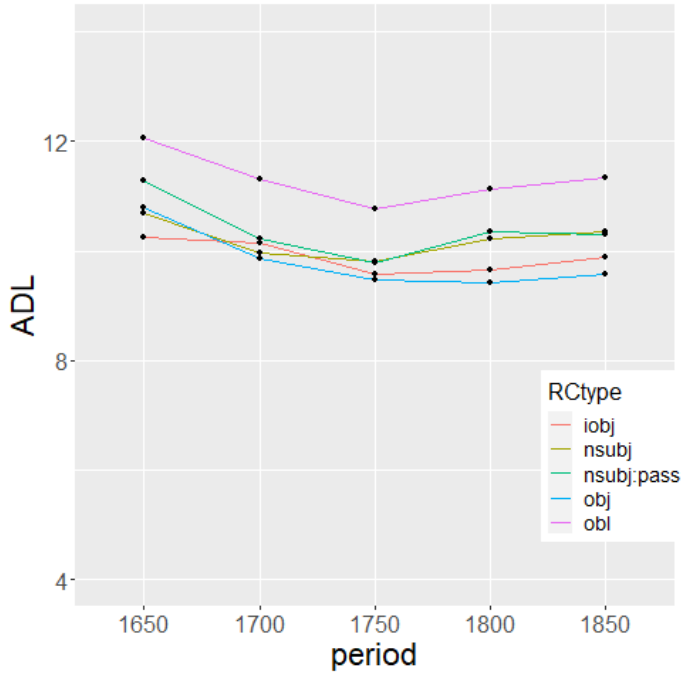


Figure 10.44: ADL per RC types per 50-year periods in scientific German (DTAW).

and a decrease in shorter ones (i.e. subject RCs). Also, we have to recognize that the ADL of RCs shows a contrary trend to the climactic trend of syntactic intricacy (cf. Chapter 9). Here, we see the exact opposite, which could represent a means of counterbalancing the abundant use of RCs overall by building shorter RCs. Regarding our H1.4b, we can confirm that RCs in scientific German create shorter DL over time, both on average and per RC type.

10.4 Summary

In the first part of the present chapter (Section 10.1), we started out by investigating the general trends in dependency length. For this, we calculated the gross average dependency length, i.e. the average across all DLs in a 50-year period (ADL) in each of the four corpora (Section 10.1.1). Our overarching hypothesis (H1.4a) was that scientific writing develops towards stronger locality, i.e. shorter ADL over time, to reduce cognitive load on working memory. Our analysis confirmed that gross average ADL indeed decreases in all four corpora. We also found that in all four corpora, sentence length (SL) seemed to be correlated with ADL, suggesting that the observed reduction in gross average ADL is driven by two main factors: (a) the prevailing SL in a time period, i.e. more short sentences lead to overall shorter ADL, and (b) the frequency distribution of short and long dependencies, i.e. the more short dependencies and the fewer long dependencies are built, the shorter is the overall ADL in a corpus.

For this reason, in the next section (10.1.2), we normalized ADL per SL. We found that only in the scientific corpora, ADL decreased diachronically independently of SL,

while in the general corpora, we did not find a chronological decrease in normalized ADL. Moreover, the chronological decrease was not found uniformly for all SL.

To find out at which SLs ADL did decrease, in Section 10.2.1, we tested three highly frequent SLs in each corpus for significant decreases in ADL over time. We found that at the highly frequent SL30, ADL showed a significant decrease in both scientific corpora. In the general corpora, ADL did not decrease significantly at any of the highly frequent SLs we tested for.

Therefore, in Section 10.2.2, we inspected SL30 more closely by calculating the ADL of each UD-relation to account for the specific factors involved in the diachronic reduction in ADL per time period. The qualitative analysis showed that the overall reduction in ADL seems to be the result of a complex interplay between the frequency of long and short dependency relations and the actual ADL of individual dependency relations. In both scientific corpora, some high-frequency, short-distance dependency relations have increased remarkably over time, while most long-distance dependency relations have either become less frequent or have not changed much in frequency over time. Our findings suggest that the overall decrease in ADL in the scientific corpora is chiefly due to a proportional increase in short-distance dependency relations and a decrease in long-distance dependency relations.

In the last part of this chapter (Section 10.3), we set out to investigate the development of the ADL of RCs in scientific writing. We found that on average, RCs become longer in ADL in both scientific corpora. When examining individual RC types separately, however, we found that all RC types separately develop toward shorter DL on average, confirming our hypothesis H1.4b. We also found that some RC types (especially oblique RCs) are longer on average than others, which suggests that the frequency of longer and shorter types of RCs seems to be the relevant factor determining the overall ADL of RCs.

Overall, this chapter has confirmed our hypothesis H1.4a that scientific writing overall develops towards stronger locality by prioritizing short syntactic dependency relations on the one hand and by reducing long dependency relations on the other. Furthermore, we were able to confirm our hypothesis H1.4b, which states that RCs in scientific writing become shorter in ADL over time. From the perspective of efficiency, we can assume that the decrease in syntactic complexity as driven by stronger locality contributes to enhanced processing ease in scientific writing overall as well as in the specific case of RCs.

Chapter 11

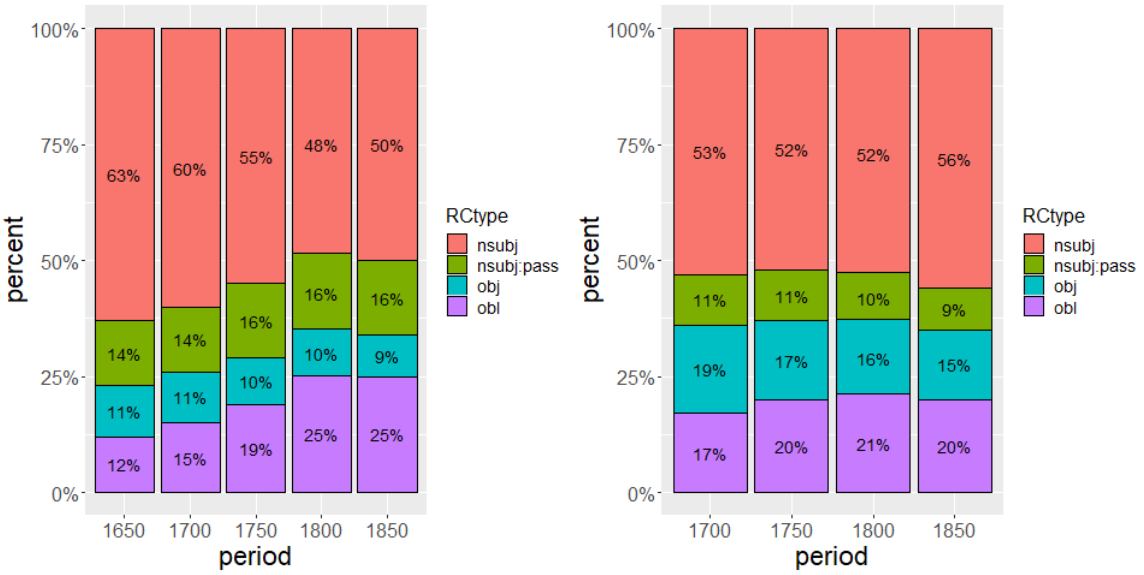
Accessibility

In the present chapter, we investigate the accessibility of RCs, the third dimension of syntactic complexity. As explained in more detail in Section 2.1.3.2, the rationale behind the concept of accessibility as an indicator of syntactic complexity is that more expected RC types (e.g. subject RCs) are easier to process than less expected ones (e.g. oblique RCs). The order in which accessibility is organized is hierarchical (from most accessible to least accessible) and was proposed by Keenan & Comrie (1977) in the Accessibility Hierarchy (AH). For our diachronic study of the development of scientific writing towards lower syntactic complexity, we thus expect that more accessible RC types should become preferred over less expected RC types lower down the AH (H1.5). In Section 11.1, we will start by testing our assumption by comparing the distributions of the different RC types in scientific writing to that in general language across the different 50-year periods. In Section 11.2, we will inspect the accessibility of RCs through the lens of the *a-score*, an aggregate measure reflecting the average accessibility of all RCs in a 50-year period. In Section 11.3, we interpret the results from our analyses and offer explanations from other complexity measures.

11.1 Frequencies of relative clause types

11.1.1 English

We start by inspecting the RC type distribution in the scientific English corpus (Figure 11.1a). In line with the AH, subject RCs are the most frequent RC type throughout all five time periods. Surprisingly, however, subject RCs decrease in proportion to other RC types: While in 1650, subject RCs were represented with a proportion of 63%, in 1850, their proportion has shrunk to 50%. The parser distinguishes between active subject RCs (`nsubj`) and passive subject RCs (`nsubj:pass`). Passive being a well-known feature of scientific writing (e.g. Biber, 1995, 2006), it is not surprising



(a) RSC **(b)** CLMET

Figure 11.1: Proportional distribution of RC types in (a) scientific (RSC) and (b) general (CLMET) English by 50-year periods.

that the proportion of passive subject RCs grows over time (cf. Example (1-a)). However, summing active and passive subject RCs, the total proportion of subject RCs shrinks from 77% to 66% (−13%) over the observed time period.

(1) a. *The platinum black under investigation was placed in the experimental tube A, and* ***the water which was pumped off*** *was collected and weighed in the U-tube E.* (Ludwig Mond, William Ramsay and John Shields, 1895) (Passive Subject RC)

b. *The action of the heat may thus be simply to increase* ***the rate at which absorption takes place.*** (ibid.) (Oblique RC)

Focusing on object RCs, we find a more expected development: the proportion of object RCs shrinks from 11% to 9%. The most surprising trend can be noted for oblique RCs (Example (1-b)), the lowest (possible in English) extraction position on the AH: the proportion of oblique RCs in scientific English rises continuously throughout the observed time period and more than doubles in 1850 compared to 1650.

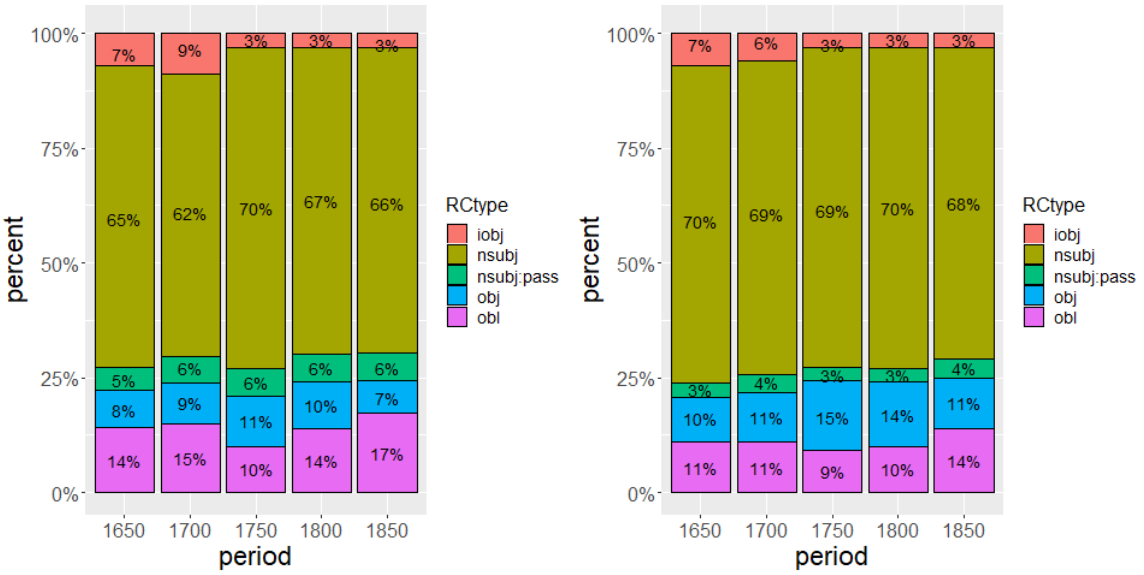
In general English (Figure 11.1b), the proportional distribution of RC types seems to be relatively stable compared to scientific English. A closer look reveals a slight but steady trend towards more active subject RCs and fewer passive subject RCs. Together, active and passive subject RCs, however, only increase by 1% over time. Interestingly, comparing the proportions in 1850 in scientific English to general English, subject RCs have a similar ratio: scientific English 66% and general English 65%. Object RCs show the strongest proportional decrease (−4%) in general English.

However, the proportion of object RCs in 1850 general English is still much higher (15%) than in scientific English (9%). As in scientific English, oblique RCs increase slightly from 17% to 20%. Note, however, that oblique RCs are already more frequent in general English in the 18th c. than in scientific English. In the 19th c., then, scientific language supersedes general English in oblique RC use (25% in scientific English and 20% in general English). The massive increase in obliques in scientific English is therefore the most noteworthy difference between the scientific and the general English corpus.

The results show that scientific English seems to shift toward a preference for two specific types of RCs, namely subject and oblique RCs, the first being highly accessible and the latter being the least accessible. The results for subject RCs are in line with observations by Biber et al. (1999) mentioning that subject RCs are most frequent in academic prose and news. They also mention that subject RCs tend to refer to given entities. When the subject gap is followed by non-subject material which provides new information, this construction meets "the informational purposes of written exposition" (Biber et al., 1999, p. 622). Although in terms of accessibility, the results for oblique RCs are surprising, in terms of their distributions across registers they are less so since Biber et al. (1999, p. 624) mention that the construction *preposition + which* (mostly representing oblique RCs, i.e. RCs with an adverbial gap) is especially common in academic prose. General English instead does not show a clear development towards a preferred type and preserves a relatively stable distribution of different RC types.

11.1.2 German

Let us now compare the RC type distributions in German (Figure 11.2). At first sight, the proportions in scientific German (Figure 11.2a) and general German (Figure 11.2b) look fairly similar: in both corpora, subject RCs occupy the largest proportion, and as in English, the scientific corpus shows a higher proportion of passive subject RCs than the general language corpus. For German, object RCs are split up into `obj` (direct object RCs) and `iobj` (indirect object RCs). Indirect object RCs decrease to a very low proportion, 3%, in 1750 in both corpora. Direct object RCs, however, show a different development. In both corpora, their proportion first increases until 1750 and then decreases.



(a) DTAW **(b)** DTAG

Figure 11.2: Proportional distribution of RC types in (a) scientific (DTAW) and (b) general (DTAG) German by 50-year periods (percentages may not total 100 due to rounding).

Overall, scientific German shows a lower proportion of direct object RCs than general German, and has a lower proportion in 1850 (7%) than in 1650 (8%). The opposite is the case for oblique RCs[1], which like in English, occupy a larger proportion of RCs in the scientific corpus than in the general language corpus. Interestingly, in 1850 both scientific and general German have an almost identical distribution of subject and non-subject RCs. Scientific German has a percentage of subject RCs of 72% and non-subject RCs of 27% (note that due to rounding the percentages do not total 100%). General German has the same percentage of subject RCs and 28% non-subject RCs. The main difference between the two corpora is the much higher proportion of oblique RCs and the much lower proportion of direct object RCs

[1]In English compared to German linguistics, the category of **oblique** nominals is somewhat ambiguous. While in German linguistics, *oblique* refers to types of cases, i.e. as an umbrella term for dative and genitive case (Wiese, 2008), in universal dependencies (UD), the oblique nominal relation is more functionally defined as "any nominal [group] (noun, pronoun, noun phrase) functioning as a non-core (oblique) argument or adjunct." This means that it functionally corresponds to an adverbial attaching to a verb, adjective or other adverb. Keenan & Comrie (1977) define it more narrowly: "'major oblique case NP [noun phrases]' (we intend here NPs that express arguments of the main predicate, as *the chest* in *John put the money in the chest* rather than ones having a more adverbial function like *Chicago* in *John lives in Chicago* or *that day* in *John left on that day*)." Thus, in Keenan & Comrie (1977)'s definition, obliques are arguments, while in UD, nominals with adjunct status are also annotated as obliques. In our analyses, we rely on the annotations in UD format. Since for German both direct and indirect objects are annotated referring to arguments such as accusative and dative objects respectively, the oblique annotation is used to refer to prepositional objects and adverbials.

in scientific German compared to general German – a pattern strongly resembling our findings for scientific English. Conducting chi-square tests to see whether the frequency distributions of RC types in scientific German differ significantly in the observed time span, as well as between two adjacent 50-year periods, we obtain p-values $< 2.2e^{-16}$ for all tests between periods, indicating that the differences in distributions are highly significant. Although not as high as in scientific English, the remarkable proportion of the difficult-to-process oblique RCs is surprising. However, unlike the distributions in English, the increase is much less pronounced. While oblique RCs in scientific English more than double in proportion and increase by more than 50%, the German obliques are relatively high in proportion from the beginning (14%) and only increase to 17%. Still, this conflicts with our assumption that difficult-to-process RC types increase proportionally.

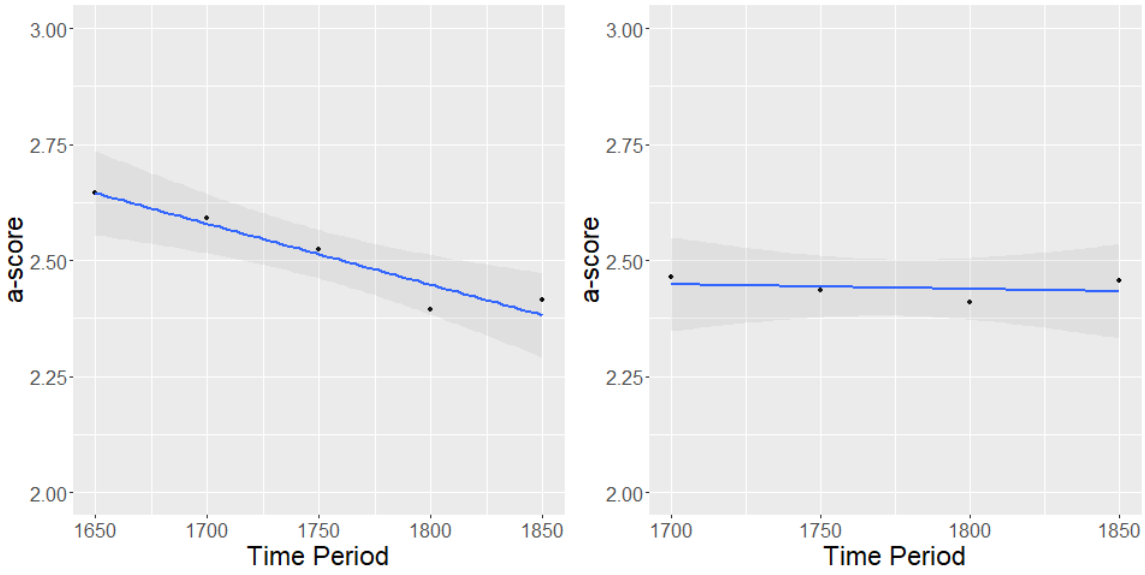
11.2 Accessibility score

The distributions of RC types can be translated to an aggregated accessibility score (a-score as described in Section 5.2.3), a measure joining all frequencies of the different RC types and reflecting the average accessibility of RCs in each corpus in each 50-year period. Passive and active subject RCs are included with the same accessibility value 1 and object RC with the value 2. Since the German UD-annotations distinguish between indirect object RCs and oblique RCs, indirect object RCs are counted with the value 3 and oblique RCs with the value 4. In English, oblique RCs are counted with the value 3.

11.2.1 English

The overall accessibility of RCs in scientific English (Figure 11.3a) shows a remarkable decrease between 1650 and 1800. The a-score in general English (Figure 11.3b) also decreases, albeit to a much lower degree. In both English corpora, the a-score increases mildly towards the end of the 19$^{\text{th}}$ c. The development in scientific English is remarkable in that the a-score starts out higher and falls lower than that of general English. The development is contrary to our hypothesis H1.5 that RCs should become more accessible and therefore less syntactically complex. Instead, the opposite seems to be the case, with the exception of the last 50-year period, where we observe a slight increase in accessibility (in both corpora). The differences between the a-scores in every two adjacent periods are all highly significant (two-sided t-test, $p < 0.005$) except for the last two periods in general English (see Table 11.1).

The relative frequencies of the different RC types (Figure 11.4) suggest that the trend of the a-score is driven by a massive decrease in subject RCs on the one hand, and a remarkable increase in oblique RCs (until the first half of the 19$^{\text{th}}$ c.) being situated lowest on the AH and therefore expected to be the hardest to process of all RC types. The relatively low a-score throughout all time periods in general English



(a) RSC **(b)** CLMET

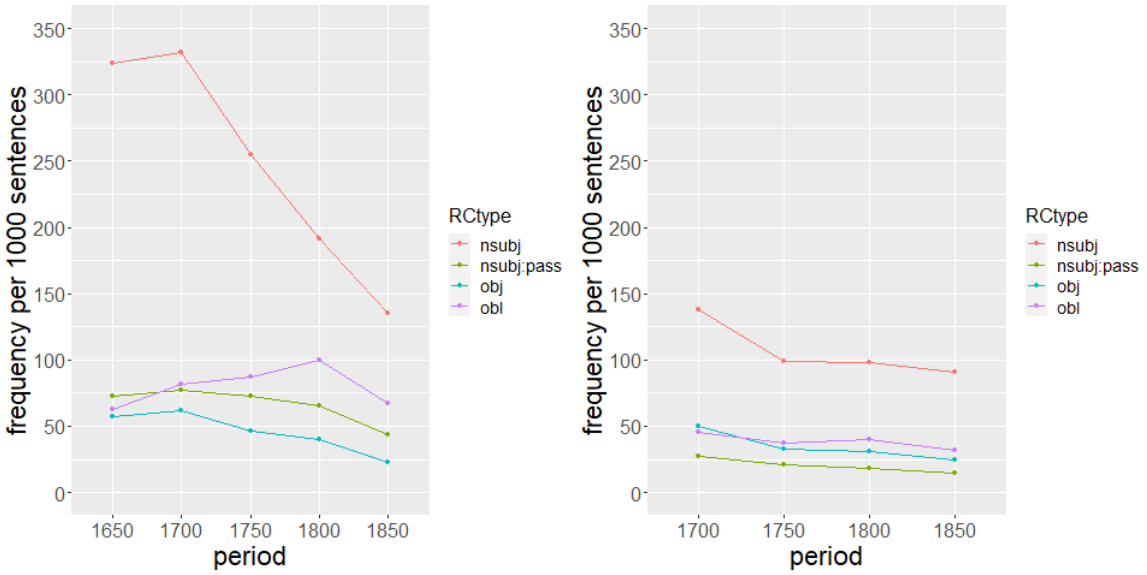
Figure 11.3: A-score in (a) scientific (RSC) and (b) general (CLMET) English by 50-year periods.

Period		**Significance $p < 0.005$**	
1	**2**	RSC	CLMET
1650	1700	**y**	–
1700	1750	**y**	**y**
1750	1800	**y**	**y**
1800	1850	**y**	**n**

Table 11.1: Significance of t-tests conducted over the a-scores between two adjacent 50-year periods in scientific English (RSC) and general English (CLMET).

can be attributed to relatively stable frequencies of the different RC types throughout the observed time span.

Summarizing our insights about accessibility in scientific English, we can report that the exact opposite of our hypothesis seems to be the case: Instead of finding higher accessibility due to stronger use of the more frequent and therefore more expected subject RCs, we found a sharp decrease in subject RCs per 1000 sentences, as well as a proportional decrease in object RCs. Both in proportion and in relative frequencies, object RCs become strongly disfavored as well, while oblique RCs, the most difficult-to-process RC types, rank second in proportional share, pulling down the accessibility score.



(a) RSC **(b)** CLMET

Figure 11.4: Relative frequencies (per 1000 sentences) of RC types in (a) scientific (RSC) and (b) general (CLMET) English by 50-year periods. All adjacent periods differ significantly from each other as determined by chi-squared tests with p-values $< 2.2e^{-16}$.

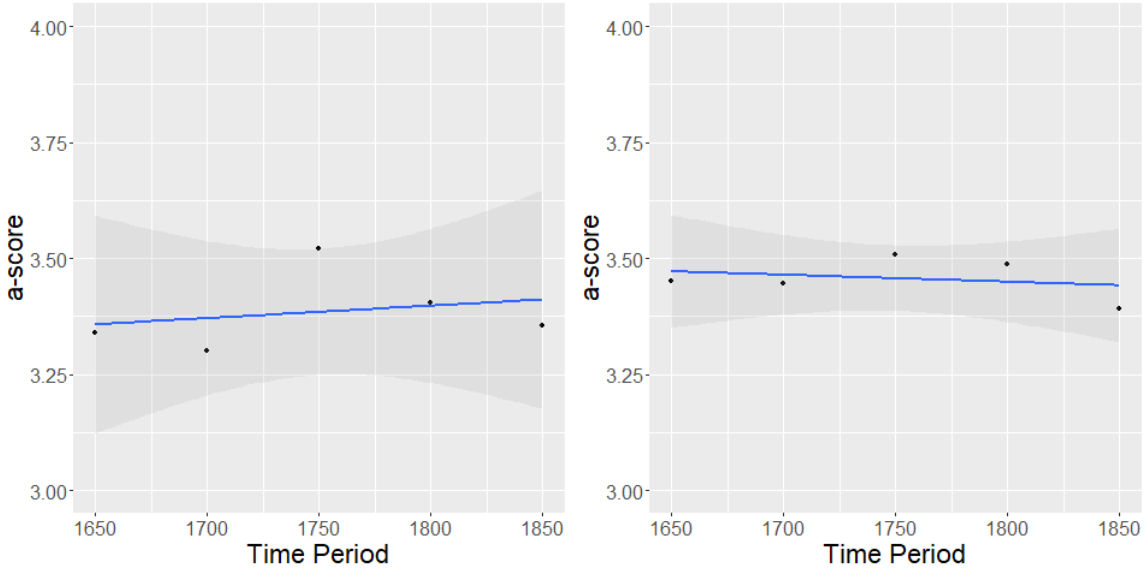
11.2.2 German

Let us now consider the a-scores for the German corpora (Figure 11.5). Note that the scale of possible values now lies between 1 and 4. This is due to the fact that the UD annotations for German RC types have a higher granularity differentiating between indirect (`iobj`) and direct object (`obj`) instead of just object RCs (`obj`) in English.

Overall, in scientific German (Figure 11.5a) we can see a mild upward trend in accessibility, while in general German, the trend is slightly downward (Figure 11.5b). The differences between the a-scores in every two adjacent periods are all highly significant (two-sided t-test, $p < 0.005$) except for the first two periods in general German (see Table 11.2). In the general German corpus, the a-scores are more stable (especially in the first two periods), while in scientific German the a-scores show stronger oscillations, starting out low, then rising towards 1750, and then falling again. This trend resembles the trend found for the RC frequencies in scientific German, giving the impression that the temporary rise in accessibility towards the end of the 18$^{\text{th}}$ c. is a result of a strong increase in subject RCs.

To verify this assumption, next, we inspect the relative frequencies of RC types per 1000 sentences (Figure 11.6). Indeed, we see that the frequency trend of subject RCs in scientific German follows the same trajectory as that of the a-scores. Figure 11.6a also shows that the increasing proportion of oblique RCs is not actually due to a frequency increase, but rather due to the fact that they stay relatively stable in

frequency over time, while all the other RC types strongly decrease in frequency toward the end of the 19th c. This shows that oblique RCs seem to be an essential type of RCs also in scientific German and it is thus worth inspecting them more closely. In general German, the frequencies of oblique RCs behave in a similar way by staying relatively stable. However, direct object RCs are fairly similar in frequency, suggesting that in both general language corpora, the choice between the RC types is more similarly distributed than in the scientific corpora. The equally high frequencies of direct object RCs and oblique RCs alongside a continuous decrease in subject RCs also seem to contribute to the slight downward trend of the a-score in general German. While the non-subject RCs are equally low and stable as in scientific German, the subject RCs show an almost linear decrease.

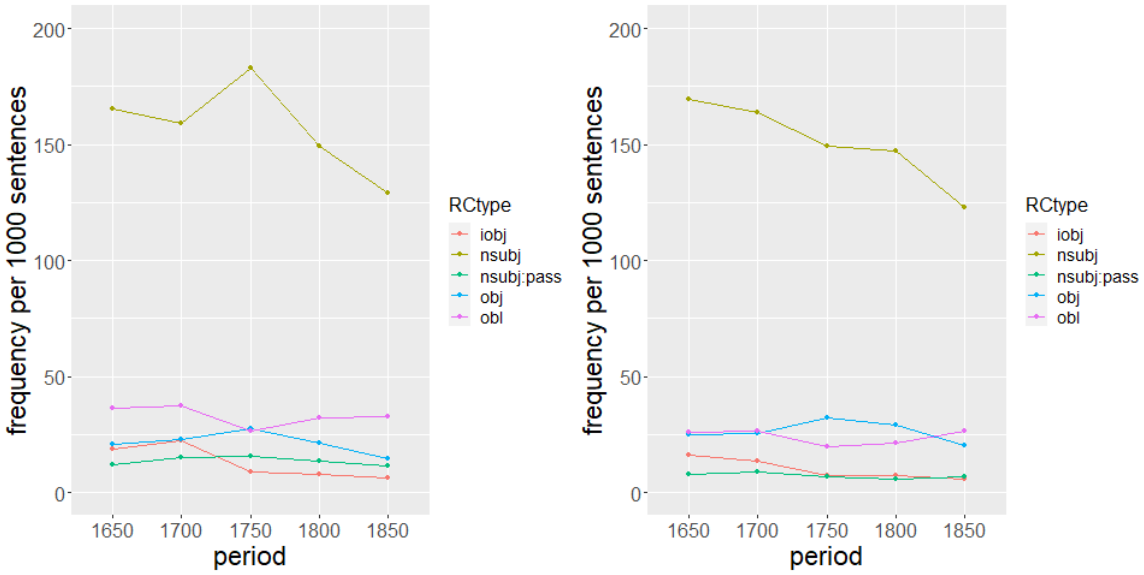


(a) DTAW **(b)** DTAG

Figure 11.5: A-score in (a) scientific (DTAW) and (b) general (DTAG) German by 50-year periods.

Period		**Significance $p < 0.005$**	
1	**2**	DTAW	DTAG
1650	1700	y	n
1700	1750	y	y
1750	1800	y	y
1800	1850	y	y

Table 11.2: Significance of differences between a-scores of two adjacent 50-year periods in scientific German (DTAW) and general German (DTAG).



(a) DTAW **(b)** DTAG

Figure 11.6: Relative frequencies (per 1000 sentences) of RC types in (a) scientific (DTAW) and (b) general (DTAG) German by 50-year periods. All adjacent periods differ significantly from each other as determined by chi-squared tests with p-values $< 2.2e^{-16}$.

11.3 Interpretation of results

The results found for accessibility do not exactly confirm our hypothesis H1.5 that scientific writing should become more accessible over time, i.e. show a preference for easy-to-process RC types such as subject RCs, and disfavor hard-to-process RC types such as oblique RCs. In the following, we will present possible explanations for the unexpected results.

11.3.1 English

The results are surprising at first regarding the underlying assumption that the frequencies of RC types should mirror the accessibility hierarchy (AH). This is only the case for subject RCs, which despite exhibiting a slight decrease, preserve their position as the most frequent RC type. Although the AH suggests that object RCs should be ranked second and oblique RCs third on the scale, the opposite is observed. Both in scientific and general English, oblique RCs are the second most frequent RC type, which is inconsistent with corpus findings from contemporary corpora and contradicts the assumption that more difficult RC types should appear less frequently than easier-to-process RC types. Thus, accessibility overall does not seem to be used for making scientific communication more efficient over time. An alternative interpretation of the extreme loss of (especially subject) RCs in scientific English as a sign

of economizing language could be the aim of reducing intricacy (as we have seen in Chapter 9) where only those RC types "survive" that cannot be replaced by other ways of encoding a message. Both subject RCs and object RCs are easily replaceable in English with alternative, more compressed clausal renderings such as indefinite participle clauses (PCs). In many cases, even more compressed structures such as nominal postmodifiers and attributive adjectives (which have been shown to increase strongly in Chapter 10, Section 10.2.2) are possible alternative strategies to convey similar, but less explicit meanings. For oblique RCs, however, the replacement is much more difficult. Oblique RCs represent adverbials modifying a verb (the embedded verb in the case of RCs). This means that RCs in oblique position add much more information to the head noun than the relatively simple subject and object RCs. Thus, a replacement by a more compressed rendering would mean too big an information loss. We can therefore assume that most replaceable RC types are increasingly rendered in alternative, shorter ways since preserving them would not contribute to efficient language use, while non-replaceable structures such as oblique RCs persist. This assumption is in line with the High-Cost/Low-Cost Heuristics by Levshina (2018, p. 53) supposing that a high processing cost also yields high (informational) benefits, while according to the Low-Cost Heuristic, low cost due to easy processing load leads to reduction of forms, even formal reduction. This reduction may manifest itself in the form of shorter clausal structures (e.g. PCs), phrase internal constructions (prepositional phrases and attributive adjectives), compounds, and ultimately the formation of entirely new terms. An essential part of scientific language is reporting on new discoveries, deriving conceptual insights from them, and finally assigning unique terms to them. At a stage where a concept is new and the community is not familiar with it, it is necessary to give an explicit description of the matter. Over time, when a concept becomes known, however, it may be enough to refer to the concept with a term that the community has agreed upon. Take for instance the development of chemical terms. Examples (2-a)–(2-d) show how *hydrogen* was described as *inflammable air*[2] before it received its Greek name "derived from the Greek 'hydro' and 'genes' meaning water forming."[3]

(2) a. *The last, indeed, sufficiently characterizes and distinguishes* ***that kind of air which takes fire****, and explodes on the approach of flame; but it might have been termed fixed with as much propriety as that to which Dr. Black and others have given that denomination, since it is originally part of some solid substance, and exists in an unelastic state, and therefore may be also called factitious.* (Joseph Priestley, Observations on different kinds of air, 1772)

[2]"This term was applied to hydrogen, H_2, once it was recognized as a distinct air; it was also used as a descriptive term for flammable gases or gas mixtures more generally. [Cavendish, Franklin, Priestley, Watt et al.]", cited from (Giunta, 2023).

[3]See The Royal Society of Chemistry (2023).

b. *The term mephitic is equally applicable to what is called fixed* ***air, to that which is inflammable****, and to many other kinds; since they are equally noxious when breathed by animals.* (ibid.)
c. *I know of only three metallic substances, namely, zinc, iron, and tin, that generate* ***inflammable air*** *by solution in acids; and those only by solution in the diluted vitriolic acid, or spirit of salt.* (Henry Cavendish, 1766)
d. *After exhausting the air from the jar the* ***hydrogen*** *was allowed to pass into and through it, and this process was repeated four times.* (W. C. Sturgis and Professor H. Marshall Ward, Soil bacillus of the type of De Bary, 1899)

The examples show that before receiving the name hydrogen, the chemical element was first described as a type of air that is inflammable (using an RC), then called inflammable air (noun modified by an attributive adjective) and finally received a unique term to denominate it. In this way, many former occasions in which subject RCs used to be necessary became obsolete and new terminology took over their place.

In the following, we will first inspect those alternative structures representing candidates to replace subject RCs, i.e. PCs, nominal modifiers, and attributive adjectives. Second, we will discuss the possible alternatives for object RCs, and third we will investigate how the difficult-to-process oblique RCs may have been able to stay efficient despite their low accessibility.

11.3.1.1 Subject relative clauses

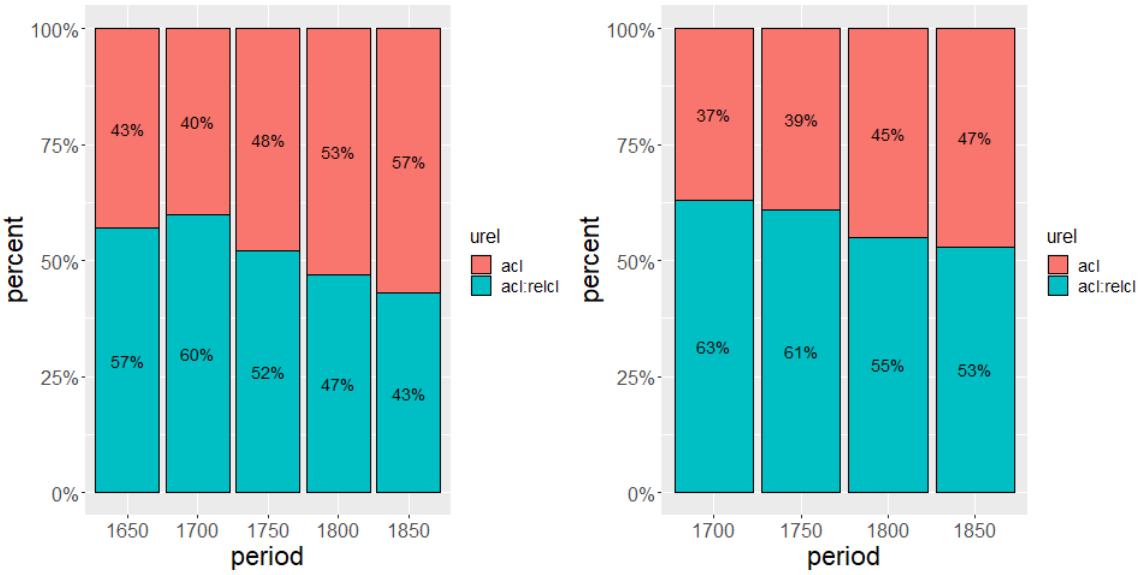
As we have seen in Figure 11.1a, in early scientific English texts (1650–1749), subject RCs were used abundantly, apparently without striving for compression, while in later texts (after 1750), RCs are used much less frequently. A possible way of rendering subject RCs in a more compressed way is using indefinite forms such as participle clauses (PCs). For instance, active subject RCs can often be paraphrased by *-ing* PCs as in Example (3):

(3) a. *It is evident that their cause is the inflection of the light* ***which comes*** *from the clouds by the sides of the hole[...].*
b. *It is evident that their cause is the inflection of the light* ***coming*** *from the clouds by the sides of the hole[...].*

Passive subject RCs can be rephrased in *-ed* PCs as in Example (4).

(4) a. *Those bodies* ***which are found*** *in certain nerves[...].*
b. *Those bodies* ***found*** *in certain nerves[...].*

To see how PCs (UD-tag: `acl`) develop in proportion over time as compared to RCs (UD-tag: `acl:relcl`) we compare the percentage distributions of PCs and RCs in the English corpora per 50-year period.

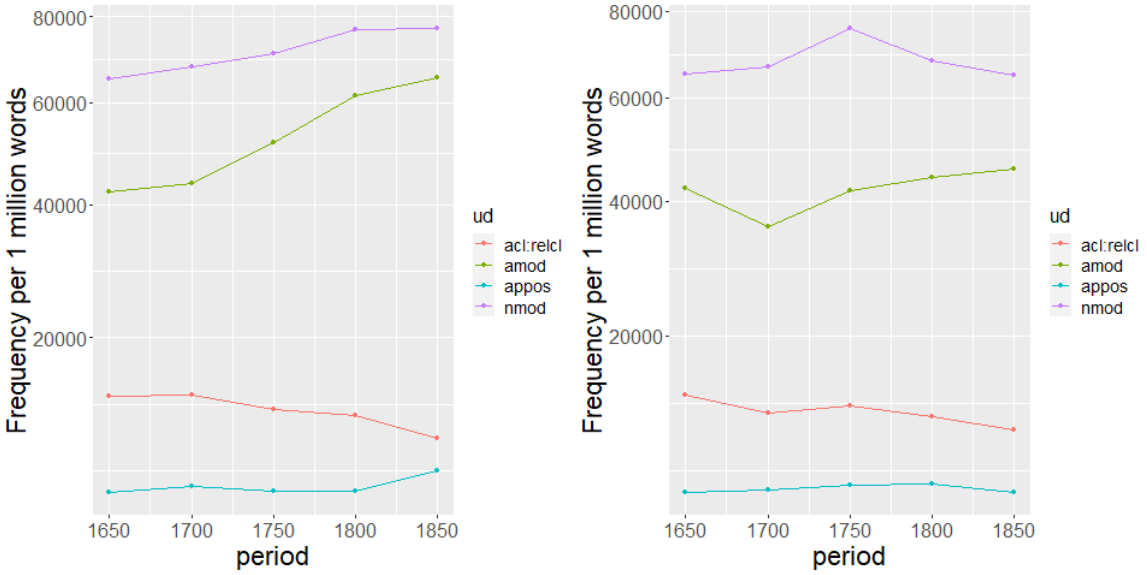


(a) RSC **(b)** CLMET

Figure 11.7: Percentage distributions of participle clauses (`acl`) and RCs (`acl:relcl`) in (a) scientific (RSC) and (b) general (CLMET) English by 50-year periods.

Figure 11.7 shows the proportions in scientific English (Figure 11.7a) and in general English (Figure 11.7b). Indeed, PCs claim an increasing proportion over time in both corpora. In scientific English, the maximal share (57%), as well as the proportional increase of PCs (+14%), is bigger than the maximal share (47%) and the proportional increase (+10%) in general English. The frequencies found for PCs in scientific and in general English furthermore differ significantly from each other (chi-squared test: p-value $< 2.2e^{-16}$); hence we may conclude that PCs are used more frequently in scientific than in general English and they seem to gradually take over an increasing part in modifying noun phrases (NPs). In terms of accessibility, this might actually mean that subject RCs do not decrease as dramatically as we would gather from our findings in Figure 11.4a observing only their full form (including the overt relativizer). Instead, they might simply have shifted towards their reduced form, exchanging part of their explicitness (through finiteness) for a higher degree of compression. Next, we will analyze the frequencies of the even more compressed forms of NP modification, namely attributive adjectives (UD-tag: `amod`), and nominal modifiers (UD-tag: `nmod`) compared to RCs (Figure 11.8). We also include appositions (UD-tag: `appos`), since they represent another compressed form of NP modification.

In scientific English (Figure 11.8a) attributive adjectives increase the most, rising by over 20,000 instances per million words in 1850 compared to 1650. Nominal modifiers increase as well, albeit less steeply. Appositions show relatively stable relative frequencies, only increasing slightly towards the end of the 19$^{\text{th}}$ c. RCs (as we have seen in our analysis on syntactic intricacy in Chapter 9) show an almost linear



(a) RSC **(b)** CLMET

Figure 11.8: Relative frequencies (per 1 million words) of attributive adjectives (`amod`), nominal modifiers (`nmod`), RCs (`acl:relcl`) and appositions (`appos`) in (a) scientific (RSC) and (b) general (CLMET) English by 50-year periods.

decrease. In general English (Figure 11.8b), the trends are less extreme. Nominal modifiers are highly frequent and increase toward a peak in 1750, while attributive adjectives first decrease and then increase. Overall, we find that attributive adjectives and nominal modifiers increase more and become more frequent in scientific English than in general English, making them strongly suspect as potentially making up for the loss in RCs. We may thus conclude that the explanation for the extreme decrease in subject RCs is very well explainable by a substantial increase in alternative NP modifying clausal constructions such as PCs, and phrasal constructions such as attributive adjectives and nominal modifiers.

11.3.1.2 Object relative clauses

We have seen that especially object RCs become extremely infrequent over time. From an accessibility point of view, this was expected, since they are lower down the AH and less easy to process than subject RCs. From an intricacy point of view, their decrease can be interpreted as a result of increasing use of their reduced variants (Example (5)):

(5) a. *The last time **that** I saw the Comet was on the 19th of October in the morning [...]*
b. *The last time _ I saw the Comet was on the 19th of October in the morning [...].*

Unfortunately, reduced RCs are extremely hard to extract with an acceptable recall, since RCs without a preceding overt relativizer are in most cases not identified as RCs. In those cases where an `acl:relcl` is annotated, precision is another problem for two reasons: reduced RCs may be confounded with oblique RCs with stranded prepositions (as in Example (6)), or they may simply be annotated erroneously. A simple evaluation of a random sample of 50 RCs without an overt relativizer resulted in only 14% precision, with many cases where an RC was annotated where there was none (compare Example (7)).

(6) *The Gold Ore* _ *we have an account* ***of*** *must be so poor as hardly to be worth taking any notice of.* (Reduced oblique RC with stranded preposition.)

(7) *As each vane passes* ***the candle it takes up heat****, and acquires extra driving energy.* (Erroneously identified as RC.)

To account for a change in the frequencies of reduced RCs automatically is therefore impossible and would require manual annotation, which unfortunately is beyond the scope of this thesis. For this reason, the possible replacement of object RCs by their reduced variants at this point must remain an assumption.

11.3.1.3 Oblique relative clauses

The rise and persistence of oblique RCs is unexpected from an accessibility point of view since they are the lowest of the AH and therefore the least easy-to-process RC type. Also, their DL is the longest of all RC types since they accommodate at least a preposition before the relativizer and at least one position for the object before the verb (see Figure 11.9).

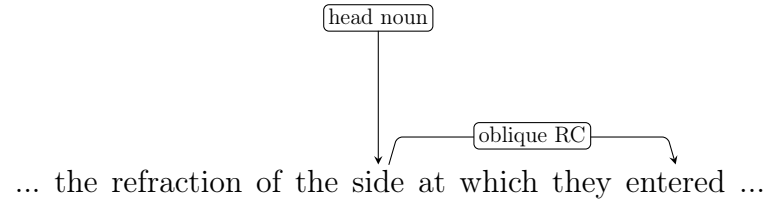


Figure 11.9: Oblique RC and intervening material stretching the DL (Henry Brougham and Charles Blagden, 1796).

Therefore, oblique RCs are inefficient in terms of two types of complexity: *accessibility* and *locality*. It is therefore extremely interesting to find out why they actually increase over time. The first idea that comes to mind is that they might be used as replacements for other even less efficient renderings. We have found in our analysis of paradigmatic richness (Chapter 6) that especially the formerly frequent pronominal adverbs (PAs) have practically disappeared from the relativizer paradigm, while the relativizer *which* became the most preferred relativizer option. PAs can be paraphrased in analytic form by the relativizer *which* + *preposition*: see Example (8).

(8) a. *[...] the Membrane immediately encompassing that skin,* ***wherein*** *the Faetus is wrapped [...].* (Philosophical Transactions, 1665–1678)
b. *[. . .] the Membrane immediately encompassing that skin,* ***in which*** *the Faetus is wrapped [. . .].* (generated alternative)

Our explanation for the loss of PA richness in the relativizer paradigm was, on one hand, that processing is easier the fewer choices we have at a given choice point, i.e. that of a relativizer introducing an RC. In our analysis of syntagmatic contexts of RCs (Chapter 7), we furthermore found that *which* increasingly occurs after pied-piped prepositions, which increases the syntagmatic predictability of RCs.

Now pulling these findings together, oblique RCs may be more complex and harder to process on the level of accessibility, while our surprisal study indirectly suggests that oblique RCs may actually be more efficient on the level of syntagmatic predictability since the constructions with the lowest surprisal are those representing oblique RCs[4]. We now want to corroborate this by extracting the different RC types with their surprisal values (Figure 11.10).

Figure 11.10 shows the distributions of the obtained surprisal values for *which* in the four different extraction positions: subject, passive subject, object, and oblique. The differences between the medians of the surprisal value distributions grow bigger over time, i.e. the median surprisal of *which* in object position is higher in 1850 (4.92) than in 1650 (4.74), while the median surprisal of *which* in oblique position is lower in 1850 (2.72) than in 1650 (3.36). The difference between the medians of object and oblique *which* in 1650 is, therefore, smaller (1.38) than in 1850 (2.2).

[4]A note on (reduced) oblique RCs with preposition stranding: The caveat with non-overt RCs in differentiating between reduced object RCs and reduced oblique RCs with stranded prepositions is also problematic if we want to explain the reasons for the increase in oblique RCs until 1800, since we have a blind spot on their reduced variants. We know from the literature (e.g. Bergh & Seppänen, 2000) that in lModE there was a prescriptive preference for overt oblique RCs with pied-piping over the use of preposition stranding. However, we are not able to determine this factor, since we are unable to detect these cases automatically. Even if the `acl:relcl` is identified correctly and annotated correctly on the embedded verb, the type of RC is not annotated, since the extraction position is annotated on the relativizer, which in this case does not exist. So, we have no automatic way of identifying the RC type of a reduced variant. The development of the two possible renderings of oblique RCs has, however, been examined in several historical linguistic corpus studies for eModE (Rydén, 1966; Ingels, 1985; Bengtsson, 1996; Lindelöf, 1997) and for lModE (Bengtsson, 1996; Van den Eynden, 1996; Johansson & Geisler, 1998; Trotta, 2000, all reviewed by Bergh & Seppänen (2000)). The reported studies unanimously come to the conclusion that not only is pied piping in RCs the prescriptively preferred structure, but also before the birth of an English standard written language, pied piping was by far the most frequent structure and the stranding option merely a "minority usage" (Bergh & Seppänen, 2000). The studies even reflect a proportional increase of pied piping (+5%) and a decrease of preposition stranding (−5%) in lModE compared to eModE. This means that on the one hand, we are at least not overlooking a high-frequency phenomenon masked by our automatic annotation. On the other hand, it means that the rise in oblique RCs with pied piping may in part be explained by a decrease in oblique RCs with stranded prepositions, since the latter have become even more infrequent than they used to be in eModE.

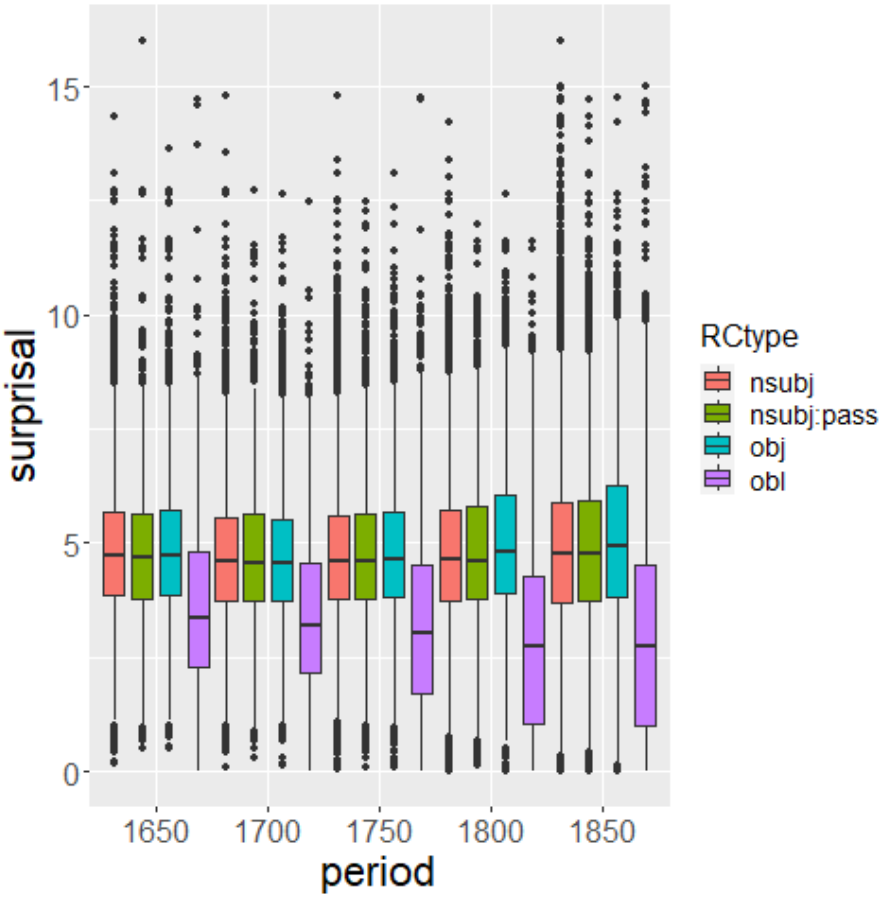


Figure 11.10: Surprisal values of *which* per RC type in scientific English (RSC).

To show that the surprisal goes down significantly over time, but without running into the problem of non-comparability between periods, we calculate the respective differences between the median surprisal of the three RC types `nsubj`, `nsubj:pass`, `obj` and `obl` (Table 11.3). In this way, we obtain comparable figures that we can use to calculate a one-sided t-test to determine the significance of the difference between the medians in one period compared to another period.

We conduct an unpaired, one-sided two-sample t-test to show that the mean difference between the differences is smaller in 1650 than that in 1850. Our *H0* is that the difference between the mean of median surprisal values in 1650 and 1850 is 0, and the *H1* is that the true difference in means between group 1650 and group 1850 is

	1650	1850
$medianSRP(\texttt{nsubj}) - medianSRP(\texttt{obl})$	1.35	1.38
$medianSRP(\texttt{nsubj:pass}) - medianSRP(\texttt{obl})$	1.33	2.04
$medianSRP(\texttt{obj}) - medianSRP(\texttt{obl})$	1.38	2.03

Table 11.3: Differences between median surprisal (SRP) of *which* in different RC types in scientific English (RSC).

less than 0. The conducted t-test yields $t = -12.88$, $df = 4$, and $p = 0.0001048$, so we can conclude that the average difference between the median surprisal for all RC types and the median surprisal of oblique RCs is significantly greater in 1850 than in 1650.

Having shown that the surprisal of oblique RCs decreases significantly over time, we can summarize that we have found two apparently opposing insights on expectation-based complexity in RCs over time: on the one hand, the *harder to process oblique RCs become more frequent over time and represent a growing proportion in the totality of RCs in scientific English.* At the same time, however, these types of RCs become increasingly predictable on the basis of their increasingly conventionalized syntagmatic contexts. As we mentioned in Section 2.1.3, we assume that in order for a language to stay efficient over time, there should be some kind of trade-off to level out increasing complexity on one level by decreasing complexity on another. On the basis of what we found for oblique RCs, we may conclude that this kind of trade-off has taken place by decreasing complexity on the level of syntagmatic predictability counteracting the increased processing effort required by processing an RC type from a very low position on the AH. As for the other RC types, we can conclude that the loss of the most accessible RC type, i.e. subject RCs, can be explained by the fact that over time, more efficient and syntactically less complex (in the sense of intricacy as well as locality) renderings of NP modification have become more preferred and can thus be assumed to have replaced the highly explicit but superfluously intricate constructions of subject RCs.

11.3.2 German

We have seen that both the percentage distributions as well as the a-scores in both German corpora behave in an expected way, i.e. expected and easy-to-process subject RCs are the most frequent RC type and take up an increasing proportion amongst RC types, while most other RC types lower down the AH become less frequent and take up lower proportional shares – with one exception, oblique RCs staying stable in frequency and increasing proportionally. The slight upward trend in accessibility in the scientific corpus generally confirms our hypothesis that scientific German RCs become more accessible and thus easier to process over time. However, neither the differences in percentages nor those in a-scores are as pronounced as those in scientific English compared to general English, suggesting that the choice of RC types has not changed in scientific German as substantially as in English. However, we have seen a general decrease in RCs; the most frequent type, subject RCs, decrease particularly substantially. At the same time, one single RC type, i.e. oblique RCs, stays remarkably stable, which seems to point to a similar irreplaceability of this RC type as in scientific English. Subject and object RCs, however, can often be replaced by alternative ways of NP modification; broadly the same mechanisms hold as those discussed above for English. In fact, in Section 10.2.2 we discovered that possible

alternative renderings such as nominal modifiers (UD-relation `nmod`) show the second biggest frequency increase of all UD-relations when comparing the last to the first 50-year period (1650 vs. 1850). In the following, we will analyze the possible replacement strategies available in German.

11.3.2.1 Subject and object relative clauses

As in English, the German scientific language can be supposed to have gone through a similar kind of development from explicit renderings of new concepts in the form of RCs to less explicit ones:

(9) a. *Dieses nasse und truckene* ***Menstruum, welches ich Alkahest genennet*** */ betreffende / so ist dasselbige nur* ***ein Erdsalz, welches so wol in forma liquida als sicca zu gebrauchen*** *[...].* (Johann R. Glauber, Philosophi & Medici Celeberrimi Opera Chymica, 1658.)

b. *Die Vegatabilien, sonderlich wenn sie etwas feucht zusammen kommen, fangen an zu gähren, und geben* ***Dünste, die man, weil sie sich leicht, und sonderlich durch die Electricität entzünden, brennbare Luft nennet****[...]. (Johann Friedrich Luz, 1784)*

The German examples (9-a) and (9-b) use RCs to give detailed additional information about the head noun, i.e. in (9-a) a substance (alcahest[5]) is described by what it is called, and how it can be used. In (9-b) *"Dünste"* (vapors) seem to be used equivalently to the English *"airs"* (nowadays gases) and the RC is used to describe their inflammable nature and their name "brennbare Luft" (*inflammable air*), which later will be called *"Wasserstoff"* (*hydrogen*).

We will now inspect several replacement strategies of subject and direct object RCs together since in German, the replacement works for both RC types in the same way. As in English, subject RCs can often be replaced by different renderings such as nominal modifiers in the form of post-modifying prepositional phrases (`nmod`) as in Example (10-b).

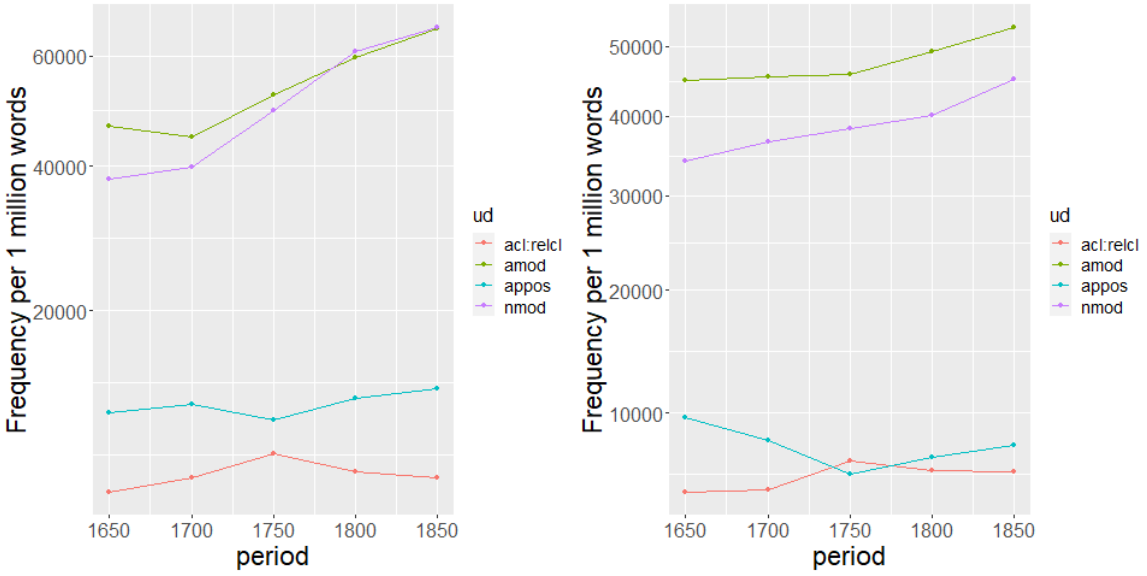
(10) a. Der Vogel, der auf dem Baum saß ...
b. Der Vogel auf dem Baum ...
c. Der auf dem Baum sitzende Vogel ...

(11) a. Die Chemikalie, die wir hinzufügten ...
b. Die hinzugefügte Chemikalie ...
c. Die Chemikalie, die wir dem Gemisch hinzufügten
d. Die dem Gemisch hinzugefügte Chemikalie

(12) a. Die Chemikalie, der wir das Wasser hinzufügten ...
b. *Die der Wasser hinzugefügte Chemikalie ...

[5]Alcahest is a hypothetical universal solvent.

Alongside simple attributive adjectives, serving as a typical means of NP premodification, another very typical alternative to German subject RCs (see Example (10)) as well as direct object RCs (Example (11)) is the participial attribute (illustrated in Examples (10-c) and (11-d). For indirect object RCs, however, this kind of paraphrase is not possible (see Example (12-b)). The participial attribute is unfortunately impossible to detect automatically in a corpus, since being a deverbal adjective it is UD-annotated as an attributive adjective (`amod`) or POS-tagged as ADJA. Supposing that the steep decrease in the subject and direct object RCs in German after 1750 may be due to the stronger use of the mentioned alternatives, we will now analyze the development of the relative frequencies (per 1 million words) of the alternative renderings, i.e. attributive adjectives (`amod`) including participial attributes, nominal modifiers (`nmod`) and appositions (`appos`), and that of RCs (`acl:relcl`) for reference (see Figure 11.11).



(a) DTAW **(b)** DTAG

Figure 11.11: Relative frequencies (per 1 million words) of attributive adjectives (`amod`), nominal modifiers (`nmod`), RCs (`acl:relcl`) and appositions (`appos`) in (a) scientific (DTAW) and (b) general (DTAG) German by 50-year periods.

We find that until 1750, RCs in scientific German, as well as attributive adjectives and nominal modifiers, become more frequent (Figure 11.11a). However, after 1750, RC frequencies decrease, while both nominal and adjectival modifiers continue to rise. Appositions show an interesting complementary trend to that of RCs, falling when RCs increase and rising when RCs decrease. Overall, the three alternative NP modifier types increase after 1650, while RCs decrease, which might point to a stepwise replacement of rendering intricate sentences in more compressed ways in scientific German. In general German, the trends are fairly similar for the alternative

nominal modifiers as well as for RCs. Note, however, that the scales for scientific and general German are slightly different since the graphs are plotted with log_2 scales for better visualization. All alternative renderings, i.e. nominal and adjectival modifiers and appositions, are more frequent in scientific German than in general German, while RCs are as well. From this, we can conclude that overall in scientific German there is a stronger tendency to richly modify NPs than there is in general German.

11.3.2.2 Oblique relative clauses

As in English, the reason for the persistence of oblique RCs may lie in the fact that they are hard to replace with other grammatical renderings, since the information loss would be too big. Furthermore, like in English, oblique RCs can be seen as analytic renderings (*preposition + d.*/welch.**) of pronominal adverbs as their corresponding synthetic relativization strategy (see Example (13)).

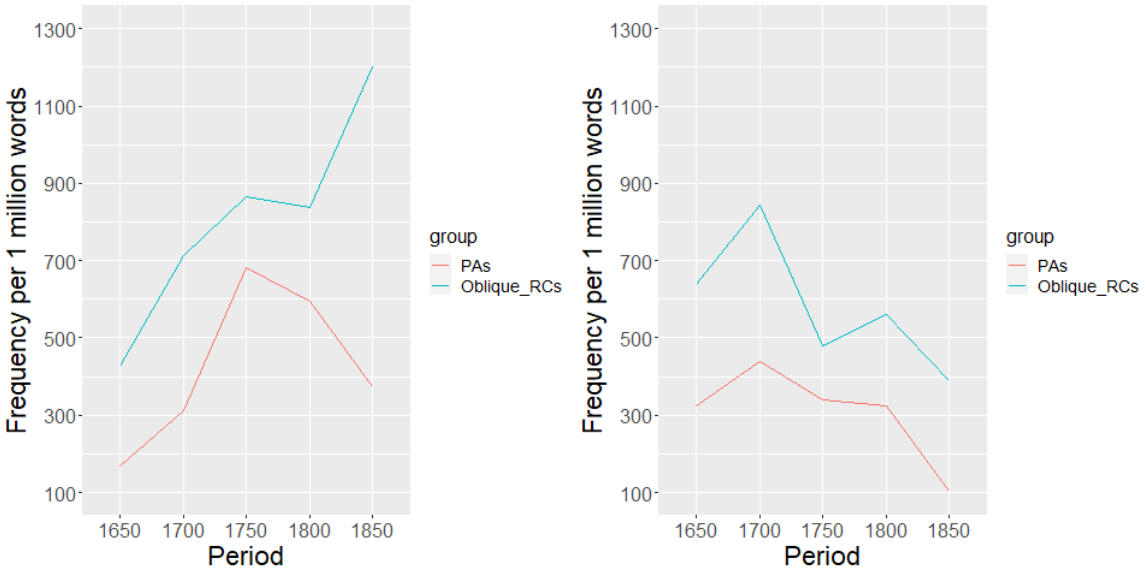
(13) a. *Dieses sind nun ohngefähr die Versuche,* ***womit*** *Harvey seine gemachte neue Entdekkung vertheidigte und bestätigte.* (Albrecht von Haller, Anfangsgründe der Phisiologie des menschlichen Körpers, 1759)

b. *Dieses sind nun ohngefähr die Versuche,* ***mit denen/welchen*** *Harvey seine gemachte neue Entdekkung vertheidigte und bestätigte.* (generated alternative)

Examples (13-a) and (13-b) show that the replacement of the PA *womit* by the analytic construction *mit denen/welchen* actually results in a more explicit version indicating the number of the head noun, *Versuche*, which a PA cannot encode. Given the fact that oblique RCs seem to represent a constant in German scientific relativization, it would be interesting to see how they have evolved compared to their synthetic, less explicit counterparts, PAs. Apart from being more explicit than PAs, oblique RCs also create longer DL than PAs since they are introduced by two words instead of one. In this line of thought, PAs would represent a more efficient variant in terms of *locality*. Also, given the assumption that scientific language should show a trend towards lower grammatical explicitness, one might expect that PAs become preferred over oblique RCs over time. Let us, therefore, compare the development of PAs and oblique RCs in terms of their frequencies per 1 million words (see Figure 11.12).

We find that the opposite of our assumption is the case: while PAs and oblique RCs show a largely parallel trend until 1800, in 1850, PAs decrease, while oblique RCs increase dramatically (Figure 11.12a[6]). We can conclude from this that PAs, despite their efficiency in terms of DL and explicitness, become strongly dispreferred in scientific German, while oblique RCs become increasingly frequent.

[6]Note that in Figure 11.12, the frequencies of oblique RCs are normalized per 1 million words and not per 1000 sentences as in Figure 11.6. The number of instances derives from the POS annotation from the corpus, i.e. *preposition + d.*/welch.** as opposed to the UD-annotation on which the numbers in Figure 11.6 are based.

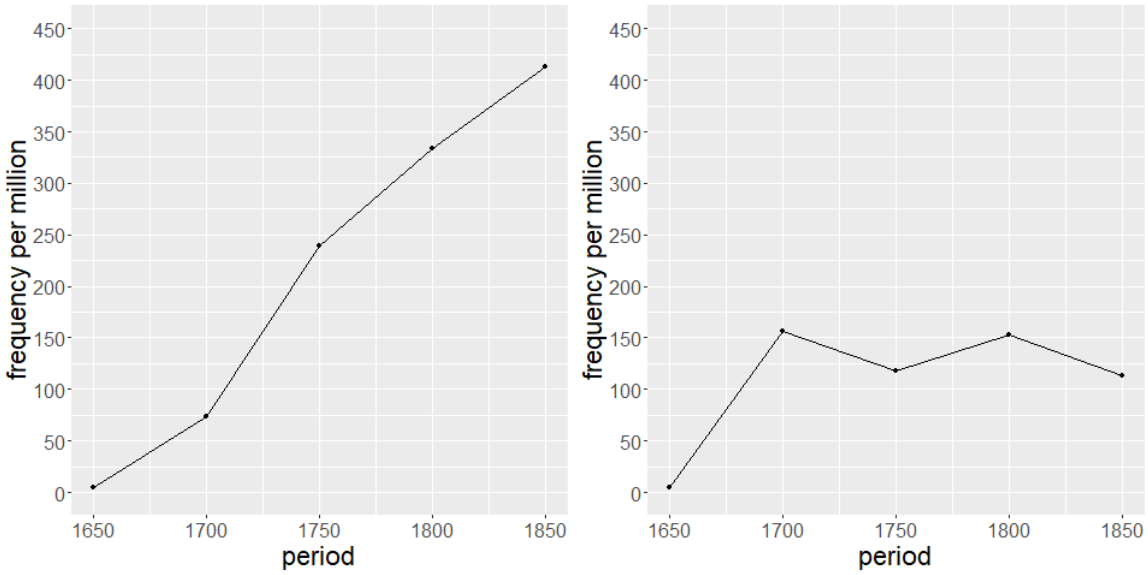


(a) DTAW **(b)** DTAG

Figure 11.12: Relative frequencies of PAs and oblique RCs per 1 million words in (a) scientific (DTAW) and (b) general (DTAG) German by 50-year periods.

In general German, both oblique RCs and PAs decrease almost simultaneously, showing that neither of the variants gains importance in relativization strategy over time in general German discourse. We would now like to explain why analytic oblique RCs become the preferred variant over their synthetic counterparts.

Our insights from our analyses on lexical complexity can offer explanations for the persistence/rise of oblique RCs in scientific German. On the one hand, for *paradigmatic richness* (Chapter 6) we have seen that the choice among the available relativizers in scientific German is relatively stable (even slightly increasing) until 1850 and then decreases drastically toward 1900 due to a major loss in PAs. In terms of expectation-based processing effort, this loss of PAs leads to a reduction of choice amongst different relativizers and thus lower processing effort in terms of entropy. In this line of thinking, on the one hand, the persistence of oblique RCs contributes to enhanced processing ease due to a complexity reduction on the lexico-grammatical level in terms of paradigmatic richness. On the other hand, syntagmatic predictability might also be involved in the preference for oblique RCs. In Chapter 7, we have shown that one of the most frequent part-of-speech trigrams preceding *welch-* was `NN PT APPR` (*noun comma preposition*), which showed a steep increase in frequency towards 1800 and dropped in frequency even more sharply in the last 50 years. While the trend of the POS trigram and the trend of oblique RCs are contrary to each other, oblique RCs are not exclusively of the form `NN PT APPR`. We therefore extract all occurrences of the POS pattern (*preposition + welch.**) (`APPR welch.*`) from the two German corpora and obtain the results shown in Figure 11.13.



(a) DTAW **(b)** DTAG

Figure 11.13: Relative frequencies (per 1 million tokens) of the pattern `preposition + welch.*` in (a) scientific (DTAW) and (b) general (DTAG) German by 50-year periods.

We can clearly see that, indeed, the pattern *preposition + welch.** increases steeply in scientific German, while in general German, it only increases between 1650 and 1700 and then stabilizes at a much lower relative frequency than in scientific German. This shows a clear preference for prepositional RCs in scientific German also on the lexico-grammatical level. As done for English above, we also would like to see whether the inefficient use of oblique RCs in terms of accessibility may be counterbalanced by a lower complexity on the level of syntagmatic predictability. To analyze the development of the surprisal of the different RC types, we extract the surprisal values of each RC type per 50-year period. The values include *welch.** and *d.**, since in German both relativizers may be used in all types of RCs (Figure 11.14).

The analysis reveals that the least frequent RC type, i.e. indirect object RCs (Example (14)), is by far the most surprising RC type in all periods. The high values are likely due to the very infrequent usage of indirect object RCs overall (compare Figure 11.2a) as well as to the fact that they often occur in the context of proper names (NE, Table 11.4). Individual proper names are generally less frequent than general nouns and thus carry a high informational load, which makes the RC less predictable in their context than in the context of a general noun (compare Examples (14-a) and (14-b).

(14) a. *Darum gebe ich diesem meinen Sali mit Fleiß den Namen Enixum nicht, auf daß die Spötter und Haderkatzen nicht eine Ursache bekommen, aus Neid und Mißgunst dawider zu lästern, sagende, daß mein Salz des Paracelsi Soll enixum nicht wäre, wie sie es gemacht mit meinem*

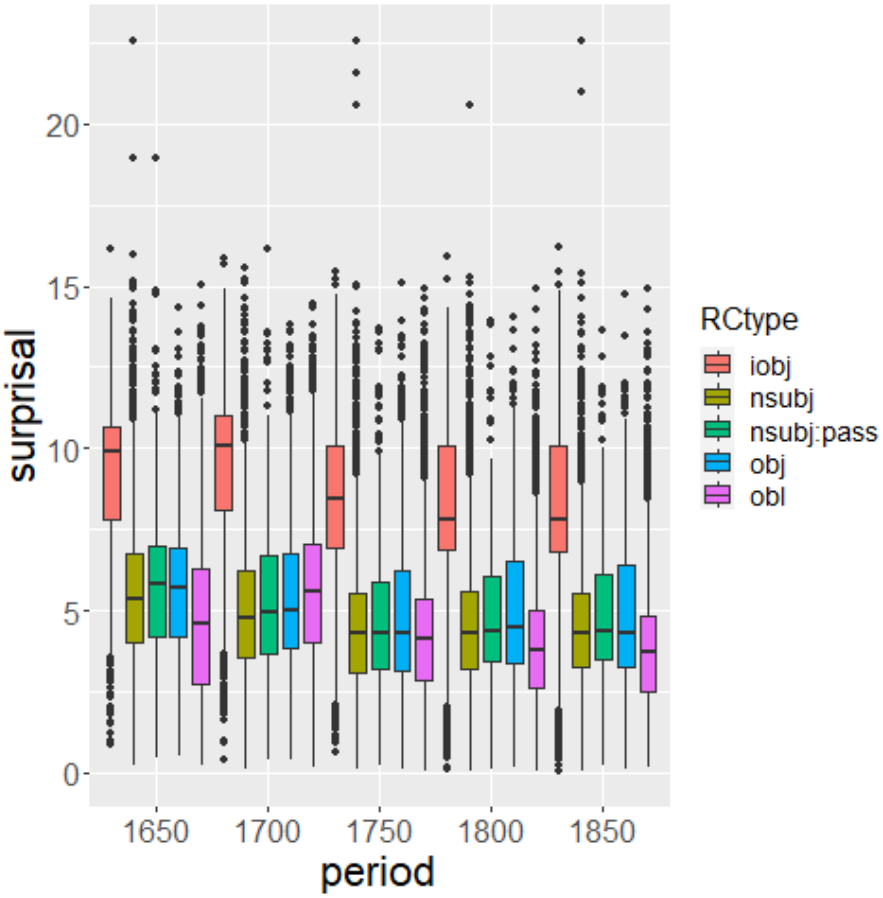


Figure 11.14: Surprisal values of *welch.** and *d.** per RC type in scientific German (DTAW).

Raw freq.	pos trigram
2125	ART NN ,
1683	ADJA NN ,
628	NN VVFIN ,
464	**NE NE ,**
355	NN VVPP ,

Table 11.4: Top five preceding POS trigram contexts of indirect object RCs in scientific German (DTAW).

Menstruo Universali, welchem *ich den Namen Alkahest gegeben.* (Glauber, Johann Rudolph: Philosophi & Medici Celeberrimi Opera Chymica. Frankfurt (Main), 1658.)

b. *Die Analogien der Tonverhältnisse mit den Abständen* ***der Planeten, denen*** *Kepler so lange und so mühsam nachspürte, blieben aber, wie mir scheint, bei dem geistreichen Forscher ganz in dem Bereich der Abstraktionen.* (A. von Humboldt, Kosmos. Entwurf einer physischen

	1650	1850
$medianSRP(\texttt{iobj}) - medianSRP(\texttt{obl})$	5.33	4.10
$medianSRP(\texttt{nsubj}) - medianSRP(\texttt{obl})$	0.74	0.55
$medianSRP(\texttt{nsubj:pass}) - medianSRP(\texttt{obl})$	1.21	0.64
$medianSRP(\texttt{obj}) - medianSRP(\texttt{obl})$	1.09	0.55

Table 11.5: Differences between median surprisal (SRP) of *welch.** and *d.** in different RC types in scientific German (DTAW).

Weltbeschreibung. 1845)

As in English, oblique RCs are the least surprising RC type in all periods (except 1700 where they are the second most surprising RC type after indirect objects). Other than in English, we cannot identify a clear trend regarding the extent to which oblique RCs are less surprising than the other RC types, since their median surprisal oscillates from extremely low compared to the other RC types (in 1650) to more surprising than the other types (except indirect object RCs, in 1700). Even without taking into account the surprisal oscillations in the intermediate periods (1700–1800), an unpaired, one-sided two-sample t-test shows that the mean difference between the differences is actually smaller in 1850 than that in 1650. The exact median values are shown in Table 11.5.

Our *H0* was that the difference between the mean of median surprisal values in 1650 and 1850 is 0 and the H1 was that the true difference in means between group 1650 and group 1850 is less than 0, which means that the mean in differences in 1850 is larger than in 1650. The conducted t-test yields $t = 0.45301$, $df = 6$, and $p = 0.6668$, so we can rule out the alternative hypothesis and conclude that the mean difference between the median surprisal for all RC types and the median surprisal of oblique RCs is *not* significantly greater in 1650 than in 1850[7]. Regardless of the surprisal trend of oblique RCs over time, we can conclude that oblique RCs are the most predictable RC type in all periods but 1700.

Raw freq.	**POS trigram**
19365	NN , APPR
3469	VVFIN , APPR
3013	NE , APPR
2561	NN , FM
2526	NN , (

Table 11.6: Top five preceding POS trigram contexts of oblique RCs in scientific German (DTAW).

[7]We also calculated the same t-test comparing mean differences between 1700 and 1850. However, the differences are not significant here, either: ($t = 0.15546$, $df = 6$, $p = 0.5592$).

The reason for their better predictability can be assumed to be the same as in English, since they are bound to occur in the context of a comma followed by a preposition ([, APPR]; compare Table 11.6)[8]. Thus, also in German, the low surprisal values for oblique RCs seem to counterbalance the fact that in terms of accessibility, oblique RCs are hard to process on the (deep) structural level, being instead easier in terms of syntagmatic predictability.

11.4 Summary

In the present chapter, we have started by looking at the proportions of the different RC types as annotated with Universal Dependencies (UD) and their proportional shares in each 50-year period (Section 11.1). Our hypothesis (H1.5) was that scientific writing should develop toward stronger accessibility by showing an increasing preference for subject RCs being the easiest-to-process RC type according to the Accessibility Hierarchy (AH) and a dispreference for RC types lower down the AH. Against our hypothesis, in scientific English, we found a strong decrease in subject RCs and an increasing preference for oblique RCs, being the RC type on the lowest position of the AH. For general English, we did not find notable differences in the proportional usage of the different RC types over time. In German, we found much more stable proportional distributions of the different RC types, with subject RCs being the far most frequent RC type. Moreover, cross-register differences in German are much less pronounced than in English, pointing to the fact that in German the RC type neither seems to be a feature of language change nor a register feature.

In Section 11.2, we calculated the aggregate a-score (average accessibility) per 50-year period. The score reflects a strong trend toward lower accessibility between periods in scientific English and relatively stable accessibility in general English. Also, in the German corpora, a relatively stable a-score reflects the fairly stable proportions of different RC types with a slight upward trend in scientific German and a slightly downward trend in general German. The relative frequencies of RC types per 50-year period showed us that across languages and across registers, subject RCs in particular, but also object RCs, decrease over time, while oblique RCs, especially in the scientific corpora, represent the most resilient RC type.

In Section 11.3, we set out to find explanations for the loss of subject and object RCs and the preservation of oblique RCs. We showed that in scientific English, subject RCs might have been gradually replaced by alternative, more compact renderings such as participle clauses as well as attributive adjectives and nominal postmodifiers and appositions. Moreover, we observed that oblique relative clauses in scientific English exhibit the least amount of surprisal compared to other types of RCs. Their surprisal further decreases over time, indicating a possible trade-off between their

[8]Both "FM" (Foreign Material) and the placeholder "(" signaling intra-sentential punctuation are erroneous POS annotations of prepositions (*von* and *mit*).

higher complexity in terms of accessibility and dependency length, and their lower complexity resulting from their improved syntagmatic predictability.

For German, we found a similar trend toward using more condensed, alternative strategies of NP modification as an explanation for the loss of subject and object RCs. The preservation of oblique RCs in German seems to follow a similar motivation as in English: although oblique RCs are lower down the AH, they are relatively frequent in German, and unlike other RC types they do not decrease in usage over time, which shows that they do not seem to be replaced by alternative renderings. At the same time, they are much more predictable than other RC types, making up for their syntactic complexity.

Overall, this chapter has shown that accessibility in English does not seem to be a feature of syntactic complexity driving language change toward stronger efficiency. Instead, we found that in fact the least accessible constructions become strongly favored over time, since they fulfill a function that cannot otherwise be replaced.

Chapter 12

Summary of Part IV

Part IV has covered the second block of our corpus analyses focusing on syntactic complexity as measured by *intricacy*, i.e. the relative frequencies of RCs per 50-year period and the relative number of RC embeddings per sentence; *locality*, i.e. generally the average dependency length (ADL) in each corpus per 50-year period and the ADL of RCs and their sub-types; and *accessibility*, i.e. the distributions of differently accessible RC types.

For ***intricacy*** (Chapter 9), we were able to confirm our hypothesis that in scientific writing RCs are used less frequently overall (H1.3a) and within individual sentences (H1.3b) over time. However, the development did not prove to be register-specific, since we found similar trends for general language. In German, we also found a decreasing use of RCs overall and within individual sentences. As expected, the development in German happened with temporal delay (H2) as compared to English (after 1750). The results have therefore shown that decreasing syntactic intricacy seems to be a cross-register and cross-lingual phenomenon of diachronic language change.

In Chapter 10, we first focused on the general development of ***locality*** in the four corpora (Section 10.1) . The analyses confirmed our hypothesis (H1.4a) that scientific writing increasingly develops toward stronger locality as shown by a decrease in ADL. As well, we could confirm that this development is specific to scientific writing (H1), and we also found the expected time-shifted development in German as compared to English (H2). Overall, the results have shown that locality seems to be a highly distinctive feature of complexity reduction in scientific writing since all our assumptions could be confirmed. Our indirect assumption that the overall locality (i.e. ADL) of a corpus could be related to the decreasing intricacy (number of RCs), however, could not be unambiguously confirmed. In Section 10.2, we found that RCs on average did not become shorter in terms of ADL. Moreover, compared to other syntactic phenomena, RCs have decreased comparatively mildly. These two factors suggest that RCs are not central to the overall reduction in ADL. Instead, ADL

seems to be strongly influenced by the increasing use of short-distance, high-frequency dependency relations such as determiners and attributive adjectives. In Section 10.3, our analysis of the development of locality within RCs showed that the ADL of RCs in scientific writing overall becomes longer, but when analyzing each RC type separately, our assumption (H1.4b) was confirmed that each RC type diachronically becomes shorter. The overall increasing ADL of all RC types together can be attributed to the fact that the generally longer RC type oblique RC becomes more frequent over time.

In our last analysis in Part IV (Chapter 11), we analyzed the development of RCs in terms of their ***accessibility***, an expectation-based measure of complexity classifying the processing difficulty of an RC hierarchically according to its type as determined by its extraction position (Accessibility Hierarchy, Keenan & Comrie, 1977). Our assumption was that harder-to-process RC types should become less frequent and easy-to-process RC types more frequent over time (H1.5). We could not confirm this assumption without reservations: Subject RCs remain the most frequent RC type in all corpora, yet they decrease remarkably in frequency, especially in scientific English but also to a lesser extent in scientific German. At the same time, the least accessible RC type, oblique RCs, become more frequent, especially in scientific English but also in German. In scientific English, the strong increase in oblique RCs and the decrease in subject RCs lead to a steep decline in overall accessibility (a-score). While these findings seem to contradict our assumption that scientific writing should become less complex and easier to process, more fine-grained analyses showed that scientific writing increasingly dispenses with superfluous and easy-to-process RC types such as subject RCs and replaces them with more compressed structures such as attributive adjectives, nominal modifiers, participle clauses, etc. The survival of the longer and less accessible oblique RCs can be explained by the fact that they are difficult to replace without a great loss of information. Moreover, processing difficulties associated with the high DL and low accessibility of oblique RCs seem to be counterbalanced at the lexico-grammatical level by lower surprisal, indicating increased predictability of the constructions given their syntagmatic contexts.

This second part of our corpus analyses has revealed a complex interplay of the three different dimensions of syntactic complexity *intricacy*, *locality*, and *accessibility*. All three measures have reflected a common pattern specific to scientific writing: the gradual abandonment of superfluous linguistic material in the form of RCs that may be replaced by more compressed renderings. At the same time, the surviving, syntactically complex constructions such as oblique RCs show such a high degree of conventionalization that they become highly predictable on a lexico-grammatical level, and therefore make low demands on processing resources. Thus, we have provided good evidence for our claim that scientific language becomes syntactically less complex and increasingly processing-friendly.

Part V

Conclusion and Outlook

Chapter 13

Conclusion

This last chapter is intended to give an overview of our main findings and draw more general conclusions for the development of scientific writing over time. In the first section (13.1) of this chapter, we discuss the main findings of this thesis by checking them against the hypotheses stated in Chapter 3. In Section 13.2, based on our findings, we aim to draw inferences regarding our fundamental hypothesis that through the use of RCs, scientific writing develops toward lower complexity and improves *efficiency* in general, and more specifically the *utility* (i.e. well-suitedness) for the purposes of scientific communication. In Section 13.3, we then discuss the limitations of the present work and how these could be improved in future work. In Section 13.4, we make an evaluation of the main linguistic and technical contributions of this work to the field and how it can be useful to the community. We close the chapter with suggestions for future work in Section 13.5.

13.1 Summary of results by hypotheses

For a better overview of the results of this thesis, we have summarized the outcomes of the different analyses in Table 13.1. The table states the hypotheses and indicates whether the described development can be observed in each subcorpus (yes/no). The color indicates whether the finding is in line with the specific hypothesis (green = yes, orange = no). The aim of our corpus investigations was to show how RCs have contributed to decreasing grammatical complexity on different linguistic levels as a register feature of scientific writing over time, as stated in our first hypothesis (**H1**). In the first block of our corpus analyses (Part III), we looked at the lexico-grammatical level of RCs, by analyzing the *paradigmatic richness* and *syntagmatic predictability* of the introductory markers of RCs, the relativizers. In the second block of our corpus analyses (Part IV), we looked at the syntactic complexity created by RCs by analyzing the degree of *syntactic intricacy* they create, the *locality* in general and within RCs,

as well as the *accessibility* of RCs. Our second hypothesis about language-specific differences (**H2**) was that we expected to find a time-shifted development in German due to its later development as an institutionalized, majority language for scientific communication in the German-speaking area. We also expected scientific German to show a rather climactic trend of grammatical complexity, i.e. to first increase and, after reaching a peak, to then decrease in grammatical complexity. For English, on the other hand, we expected to find a linear trend toward lower complexity.

			English		German	
	Hypothesis		Scientific	General	Scientific	General
LEXICO-GRAMMATICAL Complexity	H1.1: Lower paradigmatic richness		Yes	No	Yes (from 1850)	Yes
	Syntagmatic predictability	H1.2a: Higher contextual predictability (all RCs)	Yes	No	Yes	Yes
		H1.2b: Higher contextual predictability (specific relativizer)	Yes	No	Yes	No
SYNTACTIC Complexity	Intricacy	H1.3 a: Fewer RCs overall	Yes	Yes	Yes (from 1750)	Yes (from 1750)
		H1.3 b: Fewer RCs per sentence	Yes	Yes	Yes	Yes
	Locality	H1.4a: Higher locality (overall)	Yes	No	Yes	No
		H1.4b: Higher locality (in RCs)	Yes	-	Yes	-
	H1.5: Higher accessibility		No	No	Yes	No

Figure 13.1: Overview of results of corpus analyses. The color indicates whether the finding is in line with the specific hypothesis (green = yes, orange = no).

13.1.1 Lexico-grammatical complexity

The first block of corpus analyses (Part III) was concerned with investigating the development of lexico-grammatical complexity in scientific writing compared to general language, on the basis of the *paradigmatic richness* of the relativizer paradigm and the *syntagmatic predictability* of RCs in general as well as that of specific relativizers.

13.1.1.1 Paradigmatic richness

Our first hypothesis (H1.1) regarding lexico-grammatical complexity refers to the paradigmatic richness of the relativizer paradigms in English and German, i.e., **Sci-**

entific writing develops towards a reduction in paradigmatic richness as expressed by lower entropy indicating the conventionalization *of* and the lower uncertainty *about* the relativizer choice (Section 3.1.1.1). We found that this hypothesis was especially true for English, since we observed a steep, linear decrease in entropy of the relativizer paradigm in scientific English and a relatively stable entropy trend in general English. The distributions of the different relativizers available in English have shown that the entropy decrease in scientific English is due to a great loss of pronominal adverbs (PAs) and the abandonment of the relativizer *that* while gradually developing a clear preference for *which*. In contrast, general English does not change much in terms of its distributional configuration of relativizers. Also for German, we could confirm our hypothesis for scientific writing, since we found first a mild increase in entropy until 1800, and then a sudden decrease in the second half of the 19th c., which is perfectly in line with our expectations of a climactic, and compared to English, time-shifted development toward lower complexity. However, we could not confirm the hypothesis that we are dealing with a register-specific trend for German, since in general German entropy decreased steadily throughout the observed 250 years. The distributions of the different relativizers in German did, however, show a register-specific difference in the gradual development of a preferred relativizer in the paradigm: in general German clearly towards a preference for *d.**, and in scientific German an almost equal choice between *welch.** and *d.**. In both German corpora, the formerly frequent PAs are reduced drastically in the second half of the 19th c., strongly contributing to the observed entropy reduction. Our findings have shown that a reduction in paradigmatic richness by reducing the uncertainty about a relativizer in scientific language indeed seems to be a mechanism of counterbalancing increasing lexico-semantic pressures.

13.1.1.2 Syntagmatic predictability

The second hypothesis regarding lexico-grammatical complexity refers to the syntagmatic predictability of RCs in contexts. We split the hypothesis into:

- **H1.2a: surprisal at the onset of RCs decreases over time,** and
- **H1.2b: surprisal of certain, preferred relativizers decreases over time indicating their higher predictability due to conventionalized contexts. Surprisal of less preferred relativizers increases over time due to less conventionalized contexts.**

For English, we could confirm H1.2a since the surprisal of RCs including all relativizers compared to the average surprisal in the scientific corpus decreased significantly over time. Also, H1.2b could be confirmed, as the surprisal of the most preferred relativizer *which* has decreased significantly as compared to other relativizers over time. In line with our register-specific hypothesis, for general English, we found that the average surprisal of RCs has increased over time and that there was no

development of one relativizer becoming significantly more predictable as compared to the others.

For German, we could also confirm H1.2a for scientific writing. The register-specific results are, however, not as conclusive as in English, since we found a decreasing trend of average surprisal for all RCs across both registers. However, it is noteworthy that the decrease was stronger for scientific German. Regarding H1.2b, we also found a decreasing surprisal trend of a specific relativizer in scientific German, i.e. a straight decrease of the surprisal of *welch.**, while in general German the surprisal values of the three relativizer types show more parallel trends and do not diverge significantly over time. In this analysis, we did not find a time-shifted trend like the one we found for H1.1. Rather, the results reflect the specific split on preference for *d.** in general and *welch.** in scientific German.

The qualitative analyses of the grammatical and lexical contexts of relative clauses in scientific and general writing in English and German have shown a preference for specific grammatical and lexical patterns in scientific writing. The preferred grammatical context for RCs introduced by *which* in scientific English is "determiner noun preposition", while in scientific German, RCs introduced by *welch.** occur increasingly as modifiers of complex noun phrases with adjectival premodification. The most frequent lexical syntagmatic contexts of RCs in scientific English also coincide with the grammatical patterns and represent expressions of manner (*the manner in*) and quantification (*, one of*). In German, the lexical contexts of *welch.** do not coincide with the preferred grammatical patterns and mostly consist of function words such as *ist es,* syntactically forming fragments of cleft constructions and semantically functioning as parts of focus clauses with a topicalized head noun. Overall, the predictability of RCs in scientific writing has increased due to the growing preference for specific lexical trigrams preceding the preferred relativizer leading to a conventionalized usage of RCs in context. Additionally, the analysis of grammatical contexts indicates that scientific German exhibits a later trend towards lower grammatical complexity compared to scientific English.

In summary, the results on syntagmatic predictability showed that in English, complexity reduction at the lexico-grammatical level can clearly be regarded as a register-specific development. In scientific German, the reduction in lexico-grammatical complexity instead seems to be restricted to one specific relativizer. Changes in relativizer choice, meanwhile, reflect a trend towards register-specific preference where particular relativizers gradually settle in specific registers.

13.1.2 Syntactic complexity

In the second block of our corpus analyses (Part IV), we focused on the developments of syntactic complexity and the influence RCs have on it.

13.1.2.1 Syntactic intricacy

We started by analyzing the diachronic development of syntactic intricacy driven by the relative frequencies of RCs in our scientific and general language corpora. Our first sub-hypothesis was that

- **H1.3a: RCs overall will become less frequent in scientific writing**

and the second sub-hypothesis was that

- **H1.3b: the number of RC within a sentence will decrease on average in scientific writing.**

For English, we were able to confirm our hypothesis in scientific writing showing a linear decrease in both overall RC frequencies, as well as the average number of RC embeddings per sentence. However, the trend did not prove to be register-specific, since in general English RCs and the average number of RC embeddings are even less frequent than in scientific English.

For German, we found the same development as in English, which confirmed both hypotheses for scientific German but not for general German. A finding unique to scientific German was the climactic development in line with our **H2**, with an initial increase in RC usage and then a decrease toward the end of the 19th c., while in general German RCs became gradually less frequent.

Overall, the reduction of syntactic intricacy does not seem to be a register-specific feature of scientific writing, since it is reduced across registers and therefore rather seems to reflect a general diachronic trend in English and German in the Late Modern Period.

13.1.2.2 Locality

We analyzed the developments in the locality of syntactic dependencies as measured by the average dependency length (ADL) in the four corpora. Our general hypothesis was

- **H1.4a: Scientific writing develops toward shorter ADL leading to greater locality overall.**

And our hypothesis for RCs in particular was

- **H1.4b: Scientific writing develops towards shorter DL within the construction of RCs, i.e. between the head noun and the embedded verb of the RC.**

For English, the results confirmed our first hypothesis H1.4a by showing that only in scientific English, changes in average dependency length were due to an actual reduction of dependency length. In general English, the reduction was merely caused by generally shorter sentence lengths. The reduction in dependency length in scientific English was specifically due to the increasingly intense use of short-distance dependency relations such as intra-clausal noun phrase modifiers and the abandonment of longer dependency relations such as inter-clausal relations over time.

For German, we found the same developments, suggesting that a trend toward greater locality indeed seems to be a cross-linguistic register feature of scientific writing. We also found the expected climactic trend of initially slightly increasing ADL followed by a trend of decreasing ADL.

For the second hypothesis (H1.4b), we obtained similar results in both languages, showing that also within RCs, locality increases by decreasing the ADL between the head noun and the embedded verb of the RC. However, over time, types of RCs creating longer dependencies (i.e. oblique RCs) become preferred, leading to an overall higher average dependency length of RCs.

Summarizing our findings from our locality study, we can confirm a general reduction in dependency length in scientific writing over time, which can be assumed to mitigate processing effort in terms of working memory.

13.1.2.3 Accessibility

In our last corpus analysis, we looked at the only measure of syntactic complexity operating on the expectation-based side of processing. Our investigations were guided by the hypothesis that

- **H1.5: in scientific writing, over time, more accessible RC types (i.e. subject RCs) will be preferred over less accessible RC types.**

For English, we could not confirm the hypothesis, especially not for scientific English, since the least accessible RC type (oblique RCs) becomes significantly more frequent over time, while the most accessible RC type (subject RCs) becomes gradually less frequent, altogether leading to a linear decrease in overall accessibility as indicated by the a-score. For general English, we did not find significant changes in the RC type distributions, leading to a relatively stable a-score indicating little change in accessibility over time. The result is therefore in line with our hypothesis implying that no change, or a change toward lower accessibility, was to be expected in general English. For German, the results were in line with our hypothesis, since scientific German showed a mild increase in accessibility as a result of a complex interplay in the proportional reconfiguration of the different RC types. In general German, on the other hand, we even found a slight downward trend in accessibility, highlighting the register-specific trend towards higher accessibility in scientific German.

We furthermore conducted qualitative analyses to explain the unexpected outcome of the accessibility study. We found that the detected decrease in accessibility could be explained by a gradual abandonment of dispensable types of RCs that can easily be replaced by more efficient renderings such as intra-phrasal noun phrase modifications. The less accessible RC types thus seem to remain due to the fact that they are necessary means of expression in scientific communication. Analyzing their efficiency in terms of surprisal, we found that less accessible types of RCs are actually easiest to process in terms of syntagmatic predictability. In this way, one level of complexity is counterbalanced by another.

13.1.3 Summary

Our analyses of the different indicators of grammatical complexity on different linguistic levels have shown that most of these levels are actually interrelated with each other. For instance, the reduction of paradigmatic richness and the convergence on a preferred relativizer, on the one hand, also affects the syntagmatic predictability of this preferred option becoming more frequent and thus more likely to occur in specific contexts, on the other hand. In terms of syntactic complexity, we found that reductions of dependency length are mostly affected by different frequency configurations of grammatical constructions, i.e. lower usage of longer syntactic constructions, such as RCs, alongside higher usage of shorter constructions, potentially replacing RCs. Ultimately, "less efficient" constructions like oblique relative clauses remain, and in doing so, they become conventionalized and hence again easier to process.

For the two languages, we have seen that a reduction in grammatical complexity could be observed for scientific writing in isolation; however, when comparing the results to the general language, we found that sometimes the trends are similar to those in scientific writing. This was especially the case for German. This observation is plausible given the fact that the German vernacular fully penetrated the entirety of scientific text production at a much later point in time than was the case for scientific English. This makes it understandable that the register formation in German does not show the same, clear-cut trends as English, where the vernacular language had already constituted the majority of scientific text production from the beginning of our observed time span.

13.2 Implications for efficiency and utility

Our motivation to conduct the corpus analyses was to show that scientific writing compared to general language becomes less complex to counterbalance the pressures deriving from the continuous lexico-semantic expansion. We assume that this counterbalance leads to higher efficiency in expert-to-expert communication, creating higher utility of linguistic utterances in the specific communicative contexts of scientific communication. Our findings have shown that especially one type of complexity, ***locality***,

is significantly improved over time in both languages. We found that in particular the increasing choice of shorter, more compressed encodings of noun phrases has contributed to a more efficient code in scientific writing across the two languages. While the preference for a more compressed noun phrase structure is a well-known feature of scientific English (e.g. Biber & Gray, 2011a,b, 2016), its relation to memory-based processing effort as approximated by average dependency length had not been investigated yet. Also for German, we know that a condensed style had increasingly become a specific feature of scientific writing (Möslein, 1974; Roelcke, 2020); however, there are neither corpus-based studies based on the period between 1650 and 1899 nor processing-related complexity measures applied to the diachronic study of scientific German to prove these claims. To assume that grammatical complexity in scientific writing has decreased over time to improve efficiency and utility for its purposes also implies that the same complexity reduction in other registers would have a different (even disadvantageous) effect. For instance, a reduction on the lexico-grammatical level, e.g. in terms of *paradigmatic richness*, may be efficient in terms of uncertainty reduction. However, lexical variation also serves to make a text more entertaining to read. In this line of thinking, a reduction of paradigmatic richness in a language, in general, would presuppose a great loss of expressivity and would most likely not serve the communicative purposes of every situation. Also, syntagmatic predictability may not be advantageous in every communicative situation. As we could see in our analyses, ***syntagmatic predictability*** was mostly driven by a high degree of conventionalization of the preceding contexts of relativizers. The conventionalization of lexico-grammatical structures may be efficient for expository texts with the purpose of efficient transmission of information, yet for less informational registers, a high degree of conventionalized structures would imply a great sacrifice of expressivity.

In contrast to that, our results suggest that a high degree of ***intricacy*** does not seem to be an advantageous feature in any register, and keeping it low is efficient in both scientific and general language both in German and in English. This is plausible because long and extremely embedded sentences are hard to read and understand. The diachronic shift specific to scientific language from high levels of intricacy toward much lower intricacy in the later periods, however, shows that apparently, the use of RCs did have a communicative purpose at some point, namely that of defining formerly undefined concepts. The abandonment of such high levels of intricacy over time thus reflects the decreasing need for explicit explanations of concepts and the establishment of more compressed and implicit renderings instead.

Also, for ***locality***, our results suggest that across registers, languages seem to aim for efficiency by avoiding extremely long syntactic dependency relations. Across registers, this is first of all achieved by simply limiting sentence length. On top of that, scientific writing seems to have developed strategies that allow it to keep dependency length short even when the entire sentence is long by making use of an especially compressed style. This shows that while locality is efficient and advantageous to any written utterance, in scientific writing the extreme need for a high informativity

within a sentence pushes toward the usage of structural compression that goes beyond the mere limitation of the information content of a sentence through sentence length.

Finally, we have seen that optimizing scientific language by improving the ***accessibility*** of RCs does not seem to be negotiable. While oblique RCs are the least accessible RC type according to the AH, their proportional share in scientific English has grown remarkably. While the proportional growth is mostly due to the disappearance of the other, replaceable RC types, conversely this means that oblique RCs seem to be irreplaceable in their function of defining complex concepts in scientific writing. As they become more prevalent in the scientific meta-register, they also become conventionalized and structurally more predictable and thus increasingly easy to process in context. This simple mechanism suggests that even apparently inefficient ways of expression can contribute to utility, in certain contexts where these ways of expression appear frequently.

13.3 Limitations of the study

While the investigations conducted within the scope of the present thesis have contributed to our understanding of grammatical complexity in scientific writing, it is important to acknowledge the limitations of the work presented in order to contextualize and interpret the findings.

First of all, conducting a contrastive study always comes with limitations regarding the comparability of the data sources. In this thesis, we have worked with four different corpora of different origins. The scientific English corpus (RSC) consists of texts from the same organization and only includes texts from the natural sciences. Moreover, the texts from the RSC belong to the special medium journal article. The general English texts (CLMET) instead cover a variety of text types (letters, treatises, theatre plays, etc.) with their specific characteristics, and the German texts cover entire books with the typical structural components such as table of contents, subject indices, author indices, geographical indices, etc. These differences between the corpora obviously limit the comparative validity of the studies. Ideally, it would therefore be fairer to compare only texts from exactly the same text type and publishing medium. However, such an entirely fair comparison is hardly ever possible in real life. Especially the contrastive study of English and German scientific texts in the observed time period would not have been possible due to the fact that journal articles only began to be written in German after 1800. During the course of our studies, we also noted that the German general language corpus is too homogeneous to be considered a general language corpus, as it consists in equal parts of only two different registers: fiction and special-purpose non-fiction. This means that the register properties of both registers are very prominent. As a result, we also found a strong trend toward the conventionalization of syntagmatic contexts due to the strong influence of special-purpose non-fictional texts. Apart from that, general language is

a fuzzy concept since every linguistic utterance belongs to a specific communicative situation and there is no such thing as a general communicative situation. In this sense, general language as a register does not exist. We used the term *general* to mean *non-scientific*, and the comparison aims to show what is typical of scientific as compared to non-scientific discourse. In future work, comparisons with more concretely defined registers would be interesting to derive specific conclusions for these registers.

Apart from comparability issues, we have to keep in mind that in this thesis we have worked with historical language data which are especially prone to errors of automatic annotation such as POS tagging and syntactic parsing. To ensure the best possible annotation quality, we developed a special preprocessing procedure, which we have described in Section 4.3.1. We are aware that also the achieved parsing accuracy is not comparable to modern English state-of-the-art parsing accuracy; however, we still believe that the achieved accuracy is good enough to quantitatively reflect diachronic trends. In fact, the results of our studies are in line with previous work on scientific English and German, which makes us confident that our data are sufficiently reliable.

On a more general note, we should mention that studying complexity in language is a task that many linguists have worked on before and still have not conclusively answered the questions of what complexity is and how it can be captured in its entirety. In the present work, we chose to analyze the microcosm of relative clauses to trace the development of grammatical complexity across two and a half centuries. The results of our studies can therefore merely be regarded as a glimpse into a tiny aspect of grammatical complexity.

13.4 Main contributions

Despite its limitations, the current study has provided significant insights into the fields of linguistic complexity and language change. While the contributions primarily pertain to linguistic aspects, we would also like to acknowledge the technical contributions of this work.

13.4.1 Linguistic contributions

We conducted an extensive diachronic investigation spanning over 250 years of scientific writing in English and German. In order to assess the register-specificity of the trends we observed, we replicated all corpus analyses on general language corpora. This contrastive analysis between languages and registers confirms that a reduction in grammatical complexity is indeed a key aspect of the formation of the scientific meta-register across languages. Our analysis also reveals that the development of the scientific meta-register in German lagged behind that of English. Our findings support the idea that the emergence of the scientific meta-register is a gradual process

that involves adapting linguistic resources to meet new communicative needs, and that unfolds over an extended period rather than occurring abruptly.

In addition to shedding light on the role of grammatical complexity in shaping the scientific meta-register, this thesis has also advanced our understanding of linguistic complexity by identifying specific lexico-grammatical units as key indicators of complexity within the microcosm of relative clauses. These units include the relativizer paradigm, the syntagmatic contexts of relativizers, syntactic relations built by relative clauses, and their syntactic extraction positions. Our findings demonstrate that language, as a complex adaptive system, leverages these units to varying degrees in order to optimize utility in specific communicative situations. We have integrated complexity measures across different linguistic levels, including lexico-grammar and syntax, and have shown that both levels evolve over time towards an optimized code for scientific communication. Our analyses have revealed that developments at one level are intertwined with adaptations at another level; for example, a loss in accessibility is a result of a reconfiguration of the relativizer paradigm and the conventionalization of syntagmatic contexts of relativizers.

By linking the various levels of linguistic complexity to the corresponding cognitive processing demands, our research has provided a novel, cognitively motivated explanation for the diachronic changes in scientific writing. This approach expands upon previous work by considering how the cognitive effort required to process different linguistic structures relates to the observed changes in grammar over time.

13.4.2 Technical contributions

For the present work, we created four UD-annotated comparable corpora. In addition to the UD-annotations, the corpora include the complexity measures used in this work, i.e. surprisal, dependency length, and sentence length. The corpora are therefore highly valuable resources for anyone interested in syntactic developments and complexity measures such as dependency length and surprisal in English and German in the Late Modern Period. Moreover, our methodology applied to building the resources as well as the methods used to interpret the data can serve as inspirations for conducting similar research on different data sets.

In the present work, we have used complexity measures such as surprisal and dependency length, which are theoretically well-founded; however, when applied to naturalistic and more specifically diachronic language data, they have to be re-evaluated and adapted to avoid biases arising from the data. When working with surprisal, a measure based on probability distributions calculated on the basis of a specific vocabulary size, we have shown that a simple comparison of surprisal values between different corpora is not advisable, and we have therefore suggested ways to work around these limitations. By carefully teasing apart dependency length and sentence length and their frequency distributions, we have shown that claims based on average values (e.g. ADL calculated over all sentences in a 50-year period) are highly

susceptible to bias created by skewed frequency distributions. We also showed that a diachronic reduction in dependency length does not happen at all sentence lengths, but rather at the most frequent ones, while staying stable at less frequent ones. The considerations made in this thesis can thus serve as valuable inspiration for anyone working with these complexity measures.

13.5 Outlook

The present work has given an insight into how the use of relative clauses in scientific English and German has evolved over time, following the assumptions that grammatical structures become less complex to counterbalance increasing complexity at the lexico-semantic level. Since the study was focused on grammatical developments, in future work it would be interesting to shed light on the lexico-semantic developments occurring in parallel.

We have seen that despite their weak accessibility and their longer ADL, oblique RCs become more favored in scientific English while becoming more predictable syntagmatically. This finding suggests that different types of complexity might counterbalance each other. In future work, it would therefore be interesting to specifically compare ADL and surprisal for specific structures over time and investigate whether the "trade-off" between high dependency length and decreasing surprisal encountered for oblique RCs also holds for other grammatical constructions.

While our study has shed light on the differences between English and German scientific writing over a 250-year time span, it is important to acknowledge the limitation that German did not have a fully developed scientific meta-register during the earlier period. This raises questions about the development of scientific German in subsequent periods after the language was fully established as a means of scientific communication. Therefore, it would be valuable to conduct a future study examining scientific German 150 years later and compare it to a general German corpus to see if the differences between the registers become more pronounced over time, as they did in the case of scientific English. In future work, it would also be interesting to replicate the analyses on two distinct registers, such as scientific texts vs. narrative fictional text, rather than a mixed-register corpus.

We started out from the assumption that scientific writing should become less grammatically complex and therefore easier to process in terms of the specific measures we considered in this thesis. It is, however, important to note that we have only looked at a limited part of what complexity is. Furthermore, we need to keep in mind that the complexity minimization on the considered levels necessarily leads to consequences on other linguistic levels, e.g. increased lexical density and syntactic compression, as reported by Biber & Gray (2011a). Since syntax seems to become less intricate and shows stronger locality on the one hand, it would be worthwhile to explore the potential processing difficulties arising from increased lexical density and

syntactic compression.

In our analysis on *locality*, we found that in general language, the diachronic reduction in ADL is achieved by simply reducing SL. Although SL reduction is well accounted for in the literature, the underlying syntactic processes are less so. In future work, it would therefore be valuable to specifically investigate the syntactic processes facilitating SL reduction which take place in non-scientific writing.

In conclusion, the study of linguistic complexity is a vast and continually evolving field, and there is still much to explore and understand. Contrastive research especially can provide valuable insights into how complexity is modulated in different languages. Furthermore, we only explored a fraction of the vast variety of complexity measures available. Thus, applying and evaluating additional measures will provide a more comprehensive understanding of complexity and how it operates across languages and sublanguages.

Zusammenfassung

Das Thema dieser Arbeit ist die Entwicklung der sprachlichen Komplexität in der Wissenschaftssprache im Englischen und Deutschen zwischen 1650 und 1900. In dieser Epoche machte die wissenschaftliche Gemeinschaft enorme Fortschritte, was zur Gründung von Institutionen wie der *Royal Society* in Großbritannien und der *Leopoldina* in Deutschland führte. Diese Institutionen hatten nicht nur Auswirkungen auf den institutionellen Rahmen, in dem Wissenschaft betrieben wurde, und die Zusammenarbeit zwischen Wissenschaftlern, sondern auch auf die Standardisierung der Vernakularsprachen als Sprachen der wissenschaftlichen Kommunikation. Zuvor war Latein die Lingua franca der Wissenschaft in Europa gewesen. Humanistischer Zeitgeist einerseits und technische Notwendigkeit andererseits führten jedoch dazu, dass wissenschaftliche Texte immer mehr der breiten Öffentlichkeit zugänglich wurden. Dadurch wuchs auch der Anteil an in den Landesprachen verfassten Wissenschaftstexte und mit ihm das Interesse an der Standardisierung der Nationalsprachen. Gleichzeitig führte die Gründung nationaler wissenschaftlicher Institutionen zu spezifischen sprachlichen Entwicklungen im wissenschaftlichen Register. Die *Royal Society* hatte beispielsweise eine klare Vorstellung davon, wie wissenschaftliche Sprache aussehen sollte, und machte Vorschläge für die stilistische Gestaltung von Wissenschaftstexten. Sie forderte z.B., dass die englische Wissenschaftssprache frei von sprachlichen Verschnörkelungen und emotionaler Sprache sein sollte. Sie sollte klar, präzise und eindeutig sein (vgl. Baugh & Cable, 1993, S. 238). Solche Anweisungen sind zwar aus moderner linguistischer Sicht eher vage und lediglich stilistischer Natur, sie spiegeln jedoch die aktive Förderung der Bildung eines sprachlich unverwechselbaren wissenschaftlichen Metaregisters als Folge des Wandels der Bedürfnisse der wissenschaftlichen Gemeinschaft wider[1]. Es kann davon ausgegangen werden, dass das in der Entstehung befindliche Metaregister aufgrund grundlegender Veränderungen in der wissenschaftlichen Methodik und der enormen Zunahme von Erfindungen und Entdeckungen einem starkem Anstieg an lexiko-semantischer Komplexität ausgesetzt war, z.B. durch die kontinuierliche Neuentstehung technischen und wissenschaftlichen Vokabulars. Aus kognitiver Sicht ist es wahrscheinlich, dass ein solcher

[1]Dieses Metaregister umfasst alle Arten von wissenschaftlichen Texten.

Anstieg der lexiko-semantischen Komplexität eine enorme Belastung für die linguistische Verarbeitung darstellt. Die Frage, die wir uns in der vorliegenden Arbeit stellen, lautet also: Wie bleiben Autoren wissenschaftlicher Sprache trotz des wachsenden externen Drucks kommunikativ effizient? Hawkins (1994, 2004, 2014) geht beispielsweise davon aus, dass Grammatik eine regulatorische Funktion im diachronen Sprachwandel spielt und zur Aufrechterhaltung der Kommunikationseffizienz beiträgt; insbesondere durch Variation in Satzstellung, Taxis und syntaktische Einbettung. Durch Variation auf diesen Ebenen soll sprachliche Komplexität ausgeglichen werden, um den kognitiven Verarbeitungsaufwand insgesamt zu reduzieren. So gehen wir in der vorliegenden Arbeit davon aus, dass über die Zeit ein Ausgleich der steigenden lexiko-semantischen Komplexität durch eine Verringerung der lexiko-grammatischen Komplexität stattfindet und so zu einem optimierten Code für die Kommunikation unter wissenschaftlichen Experten führt. Auch die die oben zitierten stilistischen Forderungen der *Royal Society* spiegeln diese grammatische Komplexitätsreduktion wider, indem sie eine deutliche Abkehr von sprachlicher Redundanz fordern. Der beschriebene sprachliche Wandel ist jedoch nicht nur aus stilistischer Perspektive plausibel sondern auch aus funktionaler Sicht, wenn man bedenkt dass in der wissenschaftlichen Gemeinschaft das gemeinsame Expertenwissen so stark wächst, dass viele explizite grammatikalische Relationen im Laufe der Zeit überflüssig werden. Man kann diese Entwicklung insbesondere im natürlichen Zyklus der Entstehung eines neuen Fachterminus beobachten: angefangen bei seiner Beschreibung mit expliziten grammatischen Mitteln bis hin zur Etablierung des jeweiligen Fachterminus wie beispielsweise im Fall der Entstehung chemischer Stoffbezeichnungen. Nach der Entdeckung eines Stoffs wird dieser zunächst anderen Wissenschaftlerinnen vorgestellt und beschrieben und erst nach und nach bildet sich die Bezeichnungen mit einem Fachterminus heraus. Das Beispiel Wasserstoff im Englischen soll dies veranschaulichen:

(1) a. *The last, indeed, sufficiently characterizes and distinguishes* ***that kind of air which takes fire****, and explodes on the approach of flame [...].* (Observations on different kinds of air, Joseph Priestley, 1772)

DE: *[...] die Art von* ***Luft, die Feuer fängt*** *und in der Nähe von Feuer explodiert*

b. *The term mephitic is equally applicable to what is called fixed* ***air, to that which is inflammable****, and to many other kinds; since they are equally noxious when breathed by animals.* (ibid.)

DE: *[...] die* ***[Luft], welche brennbar ist*** *[...]*

c. *I know of only three metallic substances, namely, zinc, iron, and tin, that generate* ***inflammable air*** *by solution in acids; and those only by solution in the diluted vitriolic acid, or spirit of salt.* (Henry Cavendish, 1766)

DE: *[...]* ***brennbare Luft*** *[...]*

d. *After exhausting the air from the jar the* ***hydrogen***[2] *was allowed to pass into and through it, and this process was repeated four times.* (W. C. Sturgis, Professor H. Marshall Ward, 1899)

DE: *Nachdem die Luft aus dem Glas entfernt wurde, konnte der* ***Wasserstoff*** *durch es hindurch strömen [...].*

Beispiel (1) veranschaulicht, wie ein neues Konzept zunächst mit Hilfe der grammatisch komplexen und expliziten Konstruktion eines Relativsatzes (RS) beschrieben wird (vgl. Beispiele (1-a) und (1-b).) Im Laufe der Zeit etabliert sich das Konzept im gemeinsamen Wissen der Fachgemeinschaft, sodass die explizite Beschreibung des Konzeptes zunehmend überflüssig wird und somit kürzere, komprimiertere Konstruktionen wie attributive Adjektive (vgl. Beispiel (1-c)) an die Stelle der unhandlichen RS Konstruktionen treten. Am Ende dieser Entwicklung steht schließlich die Prägung eines neuen Fachterminus (Hydrogen, Beispiel (1-d)). Dies zeigt, dass RS besonders interessante Konstruktionen darstellen, um den Ausgleich lexiko-semantischer Expansion durch eine Verringerung der grammatikalischen Komplexität zu beobachten.

Um grammatische Komplexität als zentrales theoretisches Konzept dieser Arbeit einzuführen und seine Verzahnung mit anderen verwandten Konzepten wie *Effizienz* und *Utilität* darzustellen, beginnen wir die Arbeit mit einem Kapitel zum theoretischen Hintergrund (*Background*). zunächst geben wir einen Überblick über vorhandene Literatur zur kommunikativen Effizienz und thematisieren den Zusammenhang zwischen sprachlicher Effizienz und Komplexität. Es wird weiterhin darauf eingegangen, wie beide Konzepte mit der Bildung des wissenschaftlichen Metaregisters verbunden sind. Hier gehen wir davon aus, dass die Mittel zur Steigerung der Effizienz in einer Sprache zu einem gewissen Teil registerspezifisch sind. Da Wissenschaftssprache einerseits registerspezifischem, v.a. lexiko-semantischem Druck ausgesetzt ist, wird grammatische Komplexität andererseits so moduliert, dass die spezifischen kommunikativen Bedürfnisse von WissenschaftlerInnen dennoch effizient erfüllt werden. Wir nennen dieses Wechselspiel der Modulierung von Komplexität auf verschiedenen linguistischen Ebenen "Utilität" (*utility*).

Da der Begriff *Komplexität* trotz zahlreicher Definitionsbemühungen nach wie vor vage ist, wird weiterhin genauer definiert, wie wir ihn in der vorliegenden Arbeit verwenden, und auf was sich im Speziellen die *grammatische Komplexität* bezieht. Es werden weiterhin die einzelnen sprachlichen Komponenten von RS identifiziert, anhand derer Komplexität in dieser Arbeit analysiert wird. Um Aussagen darüber treffen zu können, inwieweit grammatische Strukturen in der Wissenschaftssprache weniger komplex und damit leichter zu verarbeiten sind, verwenden wir spezielle Komplexitätsmaße, von denen empirisch nachgewiesen wurden, dass sie mit dem kognitiven Aufwand bei der Satzverarbeitung in Verbindung stehen. Dieser kogni-

[2] *"inflammable air: This term was applied to hydrogen, H_2, once it was recognized as a distinct air; it was also used as a descriptive term for flammable gases or gas mixtures more generally. [Cavendish, Franklin, Priestley, Watt et al.]" zitiert von Giunta (2023).*

tive Verarbeitungsaufwand kann im Wesentlichen in zwei Typen unterteilt werden: den gedächtnisbasierten (memory-based) und den erwartungsbasierten (expectation-based) Verarbeitungsaufwand. Jedes der verwendeten Komplexitätsmaße kann außerdem einer strukturellen sprachlichen Ebene (d.h. Lexik und Syntax) zugeordnet werden. Zum Beispiel kann der Grad der syntaktischen Komplexität eines Satzes anhand seiner *syntaktischen Verflochtenheit* (*syntactic intricacy*) bemessen werden, d.h. der Länge und Tiefe der taktischen Strukturen, in denen Teilsätze zu Satzgefügen verbunden werden (Halliday & Webster, 2004, S. 33). *Syntaktische Verflochtenheit* kann somit unter anderem durch die optionale Verwendung von RS als sprachlich redundantes Material moduliert werden (wenn andere kürzere Kodierungen sie ersetzen können). So können wir einen Teil dieser syntaktischen Verflochtenheit anhand der relativen Häufigkeit von RS ermitteln. Auch RS selbst können auf mehr oder weniger komplexe Weise konstruiert werden, indem längere oder kürzere syntaktische Dependenzbeziehungen zwischen dem Antezedens des RS und seinem eingebetteten Verb gebildet werden. Diese Art von syntaktischer Komplexität (bekannt als *locality*, oder *Lokalität*) wird im Allgemeinen mit der kognitiven Verarbeitungsschwierigkeit des Arbeitsgedächtnisses in Verbindung gebracht, die an der Verarbeitung syntaktischer Relationen beteiligt ist (Gibson, 2000). Neben dem Arbeitsgedächtnis scheint jedoch auch die Erwartung eine entscheidende Rolle bei der Verarbeitung von RS zu spielen. RS können beispielsweise hinsichtlich ihrer *Zugänglichkeit* moduliert werden, was mit der syntaktischen Extraktionsposition des RS zusammenhängt, an der die Relativierung stattfindet. Laut der Zugänglichkeitspyramide von Keenan & Comrie (1977) sind Subjekt RS einfacher zu verstehen als Objekt RS, da für erstere eine höhere Erwartbarkeit besteht.

Neben der syntaktischen Komplexität, die durch und innerhalb von RS entsteht, können RS auch unterschiedliche Komplexitätsgrade auf lexiko-grammatischer Ebene aufweisen. Im Englischen und im Deutschen existieren beispielsweise eine Vielzahl an Relativpronomina, welche in ihrer Gesamtheit ein Paradigma bilden. Wir nennen den Grad der Variabilität (d.h. Frequenz und Wahrscheinlichkeitsverteilungen) des Relativpronomenparadigmas *paradigmatischen Reichtum* (*paradigmatic richness*). Um den komplexitätsbedingten Verarbeitungsaufwand von *paradigmatischem Reichtum* zu ermitteln, verwenden wir Shannon Entropie (Shannon, 1948). Shannon Entropy ist ein informationstheoretisches Maß, das auf Wahrscheinlichkeitsverteilungen verschiedener (lexikalischer) Optionen an einem bestimmten Punkt (im Satz) basiert. Sie stellt die die Unsicherheit über ein bevorstehendes Wort dar und hat sich in empirischen Studien als Indikator für Satzverarbeitungsschwierigkeit erwiesen (Genzel & Charniak, 2002). Auf das Relativpronomenparadigma angewendet, spiegelt die Entropie die Unsicherheit über die Wahl eines bestimmten Relativpronomens wider. Wenn alle Relativpronomen dieselbe Wahrscheinlichkeit haben, ist die Entropie am höchsten. Je stärker die Wahrscheinlichkeiten in Richtung eines bevorzugten Relativpronomen verzerrt sind, desto geringer ist die Entropie oder die Unsicherheit bei der Wahl des Relativpronomens.

Außerdem kann die lexiko-grammatische Komplexität von RS durch ihre *Vorhersagbarkeit* (*syntagmatic predictability*) in ihrem syntagmatischen Kontext beeinflusst werden. Das bedeutet, dass RS in stark konventionalisierten Kontexten (Beispiel (2-a)) einfacher zu verarbeiten sind als solche in eher atypischen Kontexten (Beispiel (2-b)), da sie in letzteren viel weniger vorhersehbar sind.

(2) a. *Die Transformation steht in Zusammenhang mit der Art und* ***Weise, in welcher*** *die Operation durchgeführt wird.*
b. *Die Transformation steht in Zusammenhang mit* ***den Hühnern, an welchen*** *die Operation durchgeführt wird.*

Wir können den Verarbeitungsaufwand an dem Punkt im Satz abschätzen, an dem wir das Relativpronomen lesen, indem wir Surprisal (basierend auf der Shannon-Entropie) berechnen. Surprisal ist ein weiteres informationstheoretisches Maß, das sich auf die Unerwartetheit eines Ereignisses (in diesem Fall ein Wort in einem bestimmten Kontext) bezieht und die Anzahl von Informationsbits darstellt, die zur Entschlüsselung des Ereignisses benötigt wird. In der vorliegenden Arbeit verwenden wir Surprisalwerte, welche in unseren linguistischen Korpusdaten mithilfe eines Trigramm-Sprachmodells annotiert wurden. Das Sprachmodell wurde auf verschiedenen Zeiträumen (50-Jahres Perioden) trainiert, um Unterschiede in der syntagmatischen Vorhersagbarkeit von Elementen zu verschiedenen Zeitpunkten zu erfassen.

Unser Ansatz zur Untersuchung der Komplexität von RS umfasst daher sowohl die *lexiko-grammatische* als auch die *syntaktische* Dimension von RS. Er basiert auf der Annahme, dass eine Verringerung der Komplexität in jeder der Dimensionen zu einer Verringerung des Verarbeitungsaufwands führt, um dem Druck entgegenzuwirken, der sich aus der erhöhten lexiko-semantischen Komplexität ergibt.

Die *lexiko-grammatische* Komplexität wird weiterhin unterteilt in *paradigmatischen Reichtum* (paradigmatic richness) und *syntagmatische Vorhersagbarkeit* (syntagmatic predictability). Die *syntaktische* Komplexität wird in drei Typen unterteilt: *syntaktische Verwobenheit* (syntactic intricacy), *Lokalität* (locality) und *Zugänglichkeit* (accessibility).

Nachdem die verwendeten Konzepte *Effizienz*, *Komplexität*, *Utilität* eingeführt wurden, geben wir einen Überblick über die historischen und sprachlichen Entwicklungen im englisch- und deutschsprachigen Raum, die mit der Bildung des wissenschaftlichen Metaregisters verbunden sind. Es werden zentrale historische Entwicklungen in Bezug auf die Institutionalisierung der Wissenschaftspraxis zwischen 1650 und 1900 diskutiert, von denen angenommen werden kann, dass sie die Entwicklung des wissenschaftlichen Metaregisters im englischen und deutschen Sprachraum beeinflusst haben. Diese sind die Institutionalisierung der Wissenschaft, die Standardisierung Vernakularsprachen und die wissenschaftliche Publikationspraxis in den Vernakularsprachen. Weiterhin wird ein Überblick über Tendenzen des Sprachwandels in dieser Zeitperiode im Allgemeinen gegeben und im Speziellen die Bildung des wis-

senschaftlichen Metaregisters in Englisch und Deutsch behandelt. Da RS das zentrale Thema der vorliegenden Arbeit sind, widmen wir einen Abschnitt den bisherigen Arbeiten zu ihrer spezifischen diachronen Entwicklung als Komplexitätsmarker sowie den Entwicklungen des Gebrauchs von Relativpronomen in der betrachteten Zeitspanne. Im letzten Teil des *Background* Kapitels werden die Hypothesen vorgestellt, auf deren Grundlage wir unsere Korpusstudien im Zusammenhang mit den fünf Dimensionen (paradigmatischer Reichtum, syntagmatische Vorhersagbarkeit, syntaktische Komplexität, Lokalität und Zugänglichkeit) der grammatikalischen Komplexität durchführen.

Im zweiten Teil der Dissertation werden die Korpora, die für unsere empirischen Studien verwendet wurden, sowie die angewandten Methoden zur Bemessung von Komplexität vorgestellt. Um die diachrone Entwicklung der grammatikalischen Komplexität im wissenschaftlichen Metaregister nachzuvollziehen, genügt es nicht ausschließlich wissenschaftliche Texte zu analysieren. Aus diesem Grund werden in der vorliegenden Arbeit alle genannten Dimensionen der grammatischen Komplexität sowohl in wissenschaftssprachlichen Korpora in beiden Sprachen als auch in zwei vergleichbaren allgemeinsprachlichen Korpora analysiert, sodass registerspezifische Entwicklungen vor dem Hintergrund eines Vergleichsobjekts erfasst werden können. Die für die Korpusstudien genutzten Korpora wurden aus bereits bestehenden Korpora erstellt und mit den notwendigen linguistischen Annotationen für die Studien weiter aufbereitet. Das englische Wissenschaftskorpus basiert auf dem *Royal Society Corpus* (RSC, Version 6.0 Open, Fischer et al., 2020) und besteht aus den *Proceedings and Transactions der Royal Society* zwischen 1665 und 1920. In unseren Korpusanalysen wird eine reduzierte Version des Korpus verwendet, die die Jahre 1665 – 1899 abdeckt. Zur Untersuchung von allgemeinsprachlichem Englisch wurde das *Corpus of Late Modern English Texts* (CLMET, Diller et al., 2011) aufbereitet. Die deutschen Korpora sind aus Texten des Deutschen Textarchivs (DTA, Geyken et al., 2018) zusammengestellt. Das DTA beinhaltet einschlägige Texte aus Wissenschaft, Gebrauchsliteratur und fiktionaler Prosa. Unser wissenschaftssprachliches Korpus (DTAW) wurde daher aus den wissenschaftlichen Texten aus der Zeit zwischen 1650 und 1899 und das allgemeinsprachliche Korpus (DTAG) aus den Texten aus Gebrauchsliteratur und fiktionaler Prosa erstellt. Da unsere Analysen auf linguistischer Annotation verschiedener Art basieren, stellen wir zunächst die Basisversionen der Korpora vor. Diese enthalten gängige linguistische Annotationen wie Lemmas und Wortarten, sowie das oben erwähnte informationstheoretische Maß Surprisal. Unsere Studien zur syntaktischen Komplexität beruhen auf der syntaktischen Annotation der Korpora mit Dependenzrelationen (Universal Dependencies). Da die automatische Erstellung syntaktischer Annotation auf historischen Sprachdaten aufgrund ihrer sprachlichen und graphischen Eigenschaften mit erheblichen Schwierigkeiten verbunden ist, wird außerdem der Prozess des syntaktischen Parsens im Detail beschrieben. Es wird vor allem auf die minutiöse Vorbereitung der Daten eingegangen, die einen reibungslosen Parsing-Prozess sowie die bestmögliche Parse-Qualität ermöglichen soll. Die durch

die Aufbereitung der Daten erlangte Qualität der syntaktischen Annotation liegt bei allen Korpora bei mindestens 80%. Schließlich werden die verschiedenen Maße der lexiko-grammatischen und syntaktischen Komplexität beschrieben und anhand von Rechenbeispielen erklärt. Zur Berechnung der lexiko-grammatischen Komplexität führen wir die informationstheoretischen Maße *Entropie* (zur Quantifizierung des *paradigmatischen Reichtums*) und *Surprisal* (zur Quantifizierung der *syntagmatischen Vorhersagbarkeit*) ein. Desweiteren werden die drei Methoden zur Bestimmung der syntak-tischen Komplexität in Bezug auf *Verwobenheit*, *Lokalität* und *Zugänglichkeit* erklärt.

Im dritten Teil der Arbeit werden schließlich die Korpusstudien zur lexiko-grammatischen Komplexität und im vierten Teil die Studien zur syntaktischen Komplexität vorgestellt. Jede Studie ist in einen makroanalytischen Teil unterteilt, in dem wir die in die beschriebenen Komplexitätsmaße verwenden, um den Grad der Komplexität in jeder Dimension zu bewerten, und einen mikroanalytischen Teil, in dem wir die sprachlichen Veränderungen, die sich auf die grammatikalische Komplexität in jeder Dimension auswirken, qualitativ untersuchen.

Teil III besteht aus den ersten beiden Korpusstudien, die sich mit der lexiko-grammatischen Komplexität befassen. Die erste Studie umfasst eine Makroanalyse, die die Entwicklung des *paradigmatischen Reichtums* des Relativpronomenparadigmas im wissenschaftlichen und allgemeinen Englisch und Deutsch untersucht. Dazu wird die Entropie des Paradigmas in fünf 50-Jahres Perioden berechnet. Die Untersuchungen beruhen auf der Hypothese, dass

- **Wissenschaftssprache über die Zeit einen niedrigeren paradigmatischen Reichtum des Relativpronomenparadigmas entwickelt, welcher anhand sinkender Entropiewerte abgelesen werden kann.**

Auf sprachlicher Ebene wird erwartet, dass sich der niedrigere *paradigmatische Reichtum* in der Konventionalisierung des Gebrauchs von Relativpronomen und einer damit einhergehenden niedrigeren Unsicherheit im Bezug auf die jeweilige Wahl des Relativpronomens widerspiegelt. Die Ergebnisse der Studie zeigen, dass sich in beiden wissenschaftlichen Korpora ein Trend in Richtung verringertem *paradigmatischem Reichtum* vollzieht, während der paradigmatische Reichtum in gemeinsprachlichem Englisch über die Zeit gleichbleibt. In den beiden deutschen Korpora hingegen wird der paradigmatische Reichtum über die Zeit geringer, wobei das gemeinsprachliche Deutsche diesen Trend sogar früher zeigt als das Wissenschaftsdeutsch. Im Wissenschaftsdeutschen steigt der paradigmatische Reichtum tatsächlich zunächst bis Anfang des 19. Jh. und fällt erst in der letzten 50-Jahres Periode erheblich. Die qualitativen Untersuchungen zeigen, dass der Verlust an paradigmatischem Reichtum im Wissenschaftsenglischen vor allem dem großen Verlust an relativisch genutzten Pronominaladverbien geschuldet ist, sowie der zunehmenden bis hin zur fast ausschließlichen Verwendung des Relativpronomens *which*. Das gemeinsprachliche Englisch hingegen behält eine relativ ausgegelichene Verteilung zwischen den beiden

Hauptrelativpronomen *which* und *that* bei, was die stabilen Entropiewerte über die Zeit erklärt. Im Deutschen ist der Abfall der Entropie am Ende des 19. Jh. ebenfalls mit einem großen Verlust an relativisch genutzten Pronominaladverbien zu erklären. Weiterhin zeigen die Verteilungen der einzelnen Relativpronomen, dass sich im Wissenschaftsdeutschen das Relativpronomen *welcher/welches/welche* zunehmend als Alternative zum Hauptrelativpronomen *der/die/das* etabliert.

Weiterhin untersuchen wir die syntagmatische Vorhersagbarkeit von RS insgesamt und die von bestimmten Relativpronomen im Speziellen über die Zeit. Die Hypothesen zu dieser Studie sind wie folgt:

- **H1.2a: Das Surprisal am Beginn eines RS wird geringer über die Zeit.** und
- **H1.2b: Das Surprisal von bestimmten, bevorzugt genutzten Relativpronomen wird geringer über die Zeit und bildet damit eine höhere Erwartbarkeit in konventionalisierten Kontexten ab. Das Surprisal von weniger bevorzugten Relativpronomen steigt, da sie zunehmend in unterschiedlichen Kontexten verwendet werden.**

Zur Überprüfung der Hypothesen untersuchen wir zunächst das Surprisal von RS, welches auf Grundlage ihrer lexikalischen Kontexte berechnet wurde (lexikalische Trigramme). Die Ergebnisse der Studie zeigen, dass ähnlich wie in der ersten Studie, RS im Wissenschaftsenglischen insgesamt vorhersagbarer werden, während sie im gemeinsprachlichen Englisch weniger vorhersagbar werden. Die Untersuchung des Surprisals einzelner Relativpronomen (*which*, *that*,Pronominaladverbien) zur Überprüfung der zweiten Hypothese zeigt, dass sich im Wissenschaftsenglischen eine zunehmend höhere Erwartbarkeit der Relativpronomen *which* verglichen mit den anderen untersuchten Relativpronomen herausbildet. Im allgemeinsprachlichen Englisch hingegen kann keine solche Entwicklung beobachtet werden. Im Wissenschaftsdeutschen kann ebenfalls eine gegenüber den andere Relativpronomen (*der/die/das* und Pronominaladverbien) eine vergleichweise höhere Erwartbarkeit des spezifischen Relativpronomens welche, welcher, welches ausgemacht werden. Im gemeinsprachlichen Deutschen ist eine solche Auseinanderentwicklung nicht der Fall, jedoch kann auch hier eine Erhöhung der Erwartbarkeit von *welche/welcher/welches* beobachtet werden. In der qualitativen Analyse der syntagmatischen Kontexte der typischen Relativpronomen im Wissenschaftsdeutsch und Englisch werden anschließend die häufigsten lexikalischen und grammatikalischen Kontexte analysiert, um zu verstehen in welchen Kontexten RS besonders konventionalisiert und dadurch erwartbarer werden. Die Ergebnisse der qualitativen Analyse der *grammatischen Kontexte* von RS zeigen, dass RS im Wissenschaftsenglischen zunehmend häufig auf das Muster *Artikel Nomen Präposition* folgen, während die häufigsten Kontexte von RS im Wissenschaftsdeutschen hauptsächlich komplexe Nominalphrasen darstellen (*Artikel Adjektiv Nomen*). Die häufigsten lexikalischen Kontexte von RS im Wissenschaftsenglischen sind adverbiale

Ausdrücke der Art und Weise (*the manner in*) und quantifizierende Ausdrücke (, one of). Im Deutschen folgen RS am häufigsten auf Muster aus Funktionswörtern, wie *ist es,*, welche syntaktisch als Fragmente von Cleft-Konstruktionen fungieren und semantisch Teile von Fokussätzen mit topikalisiertem Bezugsnomen darstellen.

In Teil IV wird die Entwicklung syntaktischer Komplexität untersucht, die durch RS entsteht und sich in diesen widerspiegelt. Zunächst analysieren wir die Frequenzentwicklung von RS in den vier Korpora, um zu verstehen, wie sich die syntaktische Verwobenheit der Wissenschaftssprache in Bezug auf den Gebrauch von RS entwickelt hat. Die Hypothesen für die Studie sind wie folgt:

- **H1.3a: RS werden insgesamt weniger frequent in der Wissenschaftssprache.**

- **H1.3b: Die durchschnittliche Anzahl der RS innerhalb eines Satzes sinkt.**

Für das Englische wurden die Hypothesen in Bezug auf die Wissenschaftssprache bestätigt, die einen linearen Rückgang sowohl der relativen Frequenzen von RS als auch der durchschnittlichen Anzahl an RS-Verschachtelungen pro Satz zeigt. Der Trend erwies sich jedoch nicht als registerspezifisch, da in allgemeinem Englisch die Frequenz der RS und die durchschnittliche Anzahl an RS-einbettungen pro Satz noch niedriger sind als in wissenschaftlichem Englisch. Auch für das Wissenschaftsdeutsch konnten die Hypothesen bestätigt werden, nicht jedoch für das gemeinsprachliche Deutsch, da auch hier die Frequenz und die durchschnittliche RS-anzahl pro Satz über die Zeit abnehmen. Für das Wissenschaftsdeutsche konnte auch die klimaktische Entwicklung mit einem anfänglichen Anstieg der RS-frequenz und einem Rückgang gegen Ende des 19. Jh. bestätigt werden, während der Rückgang der RS-frequenzen im allgemeinsprachlichen Deutsch linear verläuft.

Die nächste Studie befasst sich mit der Entwicklung der Lokalität in den vier Korpora. Die Analysen sind in zwei Teile aufgeteilt. Die Hypothesen sind wie folgt:

- **H1.4a: Wissenschaftssprache zeigt eine generelle Entwicklung zu verstärkter Lokalität in Form von kürzeren syntaktischen Dependenzbeziehungen.**

- **H1.4b: Wissenschaftssprache zeigt eine Entwicklung zu verstärkter Lokalität in Form von kürzeren syntaktischen Dependenzbeziehungen in Bezug auf RS, indem die durchschnittliche Distanz zwischen Antezedens und RS-verb kürzer wird.**

Im ersten Teil der Studie wurde die allgemeine Entwicklung der durchschnittlichen Dependenzlänge zwischen einem syntaktischen Kopf und seinem Dependenten in den vier Korpora gemessen. Hier wurde der Einfluss der Satzlänge und die

Verteilung von kurzen und langen Dependenzbeziehungen (z.B. solche, die durch RS entstehen) untersucht. Für das Englische bestätigten die Ergebnisse unsere erste Hypothese H1.4a, indem sie zeigten, dass nur im wissenschaftlichen Englisch die Veränderungen in der durchschnittlichen Länge der Dependenzen auf eine tatsächliche Verkürzung der Länge der Dependenzen zurückzuführen waren. Im allgemeinen Englisch wurde die Verkürzung lediglich durch allgemein kürzere Satzlängen verursacht. Die Verkürzung der Dependenzlänge war insbesondere auf die immer intensivere Verwendung kurzer Dependenzrelationen wie intraphrasale Nominalphrasenmodifikation und den Verzicht auf längere Dependenzrelationen wie Relationen zwischen Nebensätzen zurückzuführen.

Für das Deutsche fanden wir dieselben Entwicklungen, was darauf hindeutet, dass ein Trend zu größerer Lokalität in der Tat ein sprachübergreifendes Registermerkmal von Wissenschaftssprache zu sein scheint. Wir fanden auch den erwarteten klimaktischen Trend einer anfänglich leicht ansteigenden durchschnittlichen Dependenzlänge, gefolgt von einem absteigenden Trend.

Im zweiten Teil der Studie wurde die spezifische Entwicklung der durchschnittlichen Dependenzlänge in RS betrachtet, um herauszufinden, ob diese im Laufe der Zeit syntaktisch weniger komplex werden (H1.4b). Für beiden Sprachen konnten ähnliche Ergebnisse beobachtet werden. Diese zeigen, dass auch innerhalb von RS die Dependenzlängen zwischen dem Antezedenten und dem eingebetteten Verb des RS abnehmen. Mit der Zeit werden jedoch Typen von RS, die längere Dependenzen erzeugen (z.B. Oblique RS), bevorzugt, was zu einer insgesamt höheren durchschnittlichen Dependenzlänge innerhalb von RS führt. Fasst man die Ergebnisse unserer Lokalitätsstudie zusammen, so lässt sich eine allgemeine Verkürzung der Dependenzlänge in wissenschaftlichen Texten im Laufe der Zeit bestätigen. Es kann davon ausgegangen werden, dass dies zu einer Verringerung des Verarbeitungsaufwandes durch das Arbeitsgedächtnis führt.

Im letzten Kapitel von Teil IV wurde die allgemeine Zugänglichkeit von RS im Sinne der Zugänglichkeitshierarchie (Keenan & Comrie, 1977) in den vier Korpora im Zeitverlauf analysiert. Hierzu wurden die Verteilungen der verschiedenen RS-typen berechnet und analysiert.

Die Zugänglichkeit von RS kann zu den erwartungsbasierten Komplexitätsmaßen zur Ermittlung der syntaktischen Komplexität gezählt werden. Die Studie basiert auf der Hypothese, dass

- **H1.5: in der Wissenschaftssprache zugänglichere RS-typen (z.B. Subjekt RS) häufiger werden und weniger zugängliche Typen (z.B. Objekt RS) seltener werden.**

Für das Englische konnte die Hypothese generell nicht bestätigt werden und insbesondere nicht für das Wissenschaftsenglische, da hier besonders die am wenigsten zugänglichen RS-typen (Oblique RS) häufiger wurden, während die zugänglichsten

RS-typen (Subject RS) immer seltener wurden. Insgesamt führt diese Entwicklung zu einer generell niedrigeren Zugänglichkeit, welche anhand des *Accessibility Score* (a-score) berechnet wurde. Für das gemeinsprachliche English konnten keine spezifischen Trends in den RS-typenverteilungen beobachtet werden. Diese stabile Entwicklung führt folglich auch zu einem stabilen a-score. Für das gemeinsprachliche Englisch entsprechen die Ergebnisse somit der Hypothese, dass hier keine Veränderungen in Bezug auf Zugänglichkeit stattgefunden haben.

Für das Deutsche konnte die Hypothese für beide Korpora bestätigt werden, da das Wissenschaftsdeutsche einen leichten Anstieg der Zugänglichkeit aufwies. Dieser Anstieg kann als das Ergebnis aus einem komplexen Wechselspiel aus der proportionalen Umverteilung der verschiedenen RS-typen gedeutet werden. Im gemeinsprachlichen Deutschen hingegen, konnte ein leichter Abwärtstrend in der Zugänglichkeit beobachtet werden, sodass für das Deutsche eine registerspezifische Entwicklung hin zu höherer Zugänglichkeit von RS in der Wissenschaftssprache festgehalten werden kann.

Um herauszufinden, wie Wissenschaftssprache trotz erhöhter Verwendung von besonders unzugänglichen RS effizient bleibt, wurde das Surprisal für die verschiedenen RS-typen berechnet. Es konnte sowohl für das Deutsche als auch besonders für das Englische gezeigt werden, dass gerade die am wenigsten zugänglichen Obliquen RS über die Zeit am erwartbarsten werden. Somit scheint die Konventionalisierung dieses RS-typs auf der lexiko-grammatischen Ebene zu niedrigerer Komplexität geführt zu haben und gleicht somit die erhöhte Komplexität auf syntaktischer Ebene aus.

Die Analysen verschiedener Indikatoren für grammatikalische Komplexität auf unterschiedlichen linguistischen Ebenen haben gezeigt, dass die meisten dieser Ebenen auf irgendeine Weise miteinander verbunden sind. So wirken sich beispielsweise die Verringerung des paradigmatischen Reichtums und die damit verbundene Präferenz eines spezifischen Relativpronomens auf die syntagmatische Vorhersagbarkeit dieses Relativpronomens aus, da dieses als bevorzugte Option häufiger auftritt und damit in bestimmten Kontexten wahrscheinlicher wird. In Bezug auf die syntaktische Komplexität haben wir festgestellt, dass die Verringerung der Dependenzlänge vor allem durch unterschiedliche Häufigkeitskonfigurationen von grammatischen Konstruktionen zustande kommt, d.h. die seltenere Verwendung längerer syntaktischer Konstruktionen, wie RS und gleichzeitig die häufigere Verwendung kürzerer Konstruktionen, die RS potentiell ersetzen. Schließlich wurde festgestellt, dass syntaktisch komplexere, also weniger effiziente Konstruktionen wie Oblique RS, über die Zeit durch verstärkte Konventionalisisierung auf lexiko-grammatischer Ebene weniger komplex werden und damit wiederum leichter zu verarbeiten sind. Für die beiden Sprachen konnte eine Reduktion der grammatikalischen Komplexität in der Wissenschaftssprache zwischen 1650 und 1900 beobachtet werden. Die Tendenzen ähneln jedoch nicht selten denen in der Allgemeinsprache, insbesondere im Deutschen. Dieses Ergebnis ist allerdings nicht überraschend angesichts der Tatsache, dass die deutsche Vernakularsprache erst sehr viel später in der wissenschaftlichen Textproduktion dominierte, als dies beim wis-

senschaftlichen Englisch der Fall war. Dies kann erklären, warum die Registerbildung im Deutschen nicht die gleichen, geradlinigen Trends aufweist wie im Englischen, wo die Vernakularsprache bereits zu Beginn der in dieser Arbeit betrachteten Zeitspanne den Großteil der wissenschaftlichen Textproduktion ausmachte.

List of Abbreviations

Abbreviation	Definition
ADL	Average Dependency Length
AE	American English
AH	Accessibility Hierarchy
avgS	Average Surprisal
BBAW	Berlin Brandenburgische Akademie der Wissenschaften
BE	British English
BS	Bad Sentences
CLMET	Corpus of Late Modern English Texts
CONCE	A Corpus of Nineteenth-Century English
DL	Dependency Length
DTA	Deutsches Textarchiv
DTAG	Deutsches Textarchiv General Corpus
DTAW	Deutsches Textarchiv Wissenschaft Corpus
DWDS	Digital Dictionary of the German Language
eModE	Early Modern English
ENHG	Early New High German
ER	Entropy Reduction
ERH	Entropy Reduction Hypothesis
fMRI	functional Magnetic Resonance Imaging
FPM	Frequency Per Million Tokens
GS	Good Sentences
IQR	Interquartile Range
IR	Industrial Revolution
KLD	Kullback-Leibler Divergence
lModE	Late Modern English
LSP	Language for Special Purposes
MHG	Middle High German
NHG	New High German
...	...

Abbreviation	**Definition**
NP	Noun Phrase
OE	Old English
OHG	Old High German
OT	Optimality Theory
PA	Pronominal Adverb
PC	Participle Clause
PDE	Present-Day English
PDG	Present-Day German
PGCH	Performance-Grammar Correspondence Hypothesis
POS	Parts-of-Speech
PP	Prepositional Phrase
PTPRS	Philosophical Transactions and Proceedings of the Royal Society
RC	Relative Clause
RS	Relativsatz
RSC	Royal Society Corpus
RT	Reading Times
SDL	Summed Dependency Length
SL	Sentence Length
SOV	Subject-Object-Verb
SRP	Surprisal
SVcomplement	Subject-Verb-Complement
SVO	Subject-Verb-Object
UD	Universal Dependencies

List of Figures

List of Tables

Bibliography

Aarts, B., López-Couso, M. J., & Méndez-Naya, B. (2012). Late modern English syntax. De Gruyter Mouton.

Adelung, J. C. (1783). *Magazin für die Deutsche Sprache* volume 1. G. Olms Verlag.

Adler, M. (2012). The plain language movement. In P. M. Tiersma, & L. M. Solan (Eds.), *The Oxford Handbook of Language and Law*. Oxford: OUP.

Admoni, W. (1972). Die Entwicklung des Ganzsatzes und seines Wortbestandes in der deutschen Literatursprache bis zum Beginn des 19. Jahrhunderts. *Studien zur Geschichte der deutschen Sprache. Berlin (Bausteine zur Geschichte des Neuhochdeutschen 49)*, (pp. 243–279).

Admoni, W. (1985). Syntax des Neuhochdeutschen seit dem 17. Jahrhundert. In W. Besch, O. Reichmann, & S. Sonderegger (Eds.), *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung* (pp. 1538–1556). volume 2.

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

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

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, *42*, 239–273.

Anttila, A., Adams, M., & Speriosu, M. (2010). The role of prosody in the English dative alternation. *Language and Cognitive Processes*, *25*, 946–981.

Atkinson, D. (1992). The evolution of medical research writing from 1735 to 1985: The case of the Edinburgh Medical Journal. *Applied Linguistics*, *13*, 337–374. doi: 10.1093/applin/13.4.337.

Atkinson, D. (1996). The Philosophical Transactions of the Royal Society of London, 1675–1975: A sociohistorical discourse analysis. *Language in Society*, *25*, 333–371.

Atkinson, D. (1998). *Scientific Discourse in Sociohistorical Context: The Philosophical Transactions of the Royal Society of London, 1675–1975*. New York: Routledge. doi: 10.4324/9781410601704.

Ball, C. N. (1996). A diachronic study of relative markers in spoken and written English. *Language Variation and Change*, *8*, 227–258. doi: 10.1017/S0954394500001150.

Barber, C. (1997). *Early Modern English*. Edinburgh University Press.

Baron, A., & 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, UK.

Bartek, B., Lewis, R. L., Vasishth, S., & Smith, M. R. (2011). In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1178–1198. doi: 10.1037/a0024194.

Baugh, A. C., & Cable, T. (1993). *A History of the English Language*. (4th ed.). London: Routledge.

Bazerman, C. (1988). *Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science*. Wisconsin: University of Wisconsin Press.

Beneš, E. (1973). Die sprachliche Kondensation im heutigen deutschen Fachstil. *Linguistische Studien III: Festgabe für Paul Grebe, Teil 1.*, (pp. 40–50).

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

Bengtsson, P. (1996). Prepositional relatives in English: A diachronic case study of three authors. Unpublished paper. Department of English, Göteborg University.

Bergh, G., & Seppänen, A. (2000). Preposition stranding with wh-relatives: A historical survey. *English Language & Linguistics*, *4*, 295–316. doi: 10.1017/S1360674300000265.

Berlin Brandenburgische Akademie (2020). DTA Textauswahl. URL: `https://www.deutschestextarchiv.de/doku/textauswahl`.

Betten, A. (1987). Deutschsprachige Prosa vom 8. bis 17. Jahrhundert: Bestandsaufnahme und Beschreibungsprobleme. In *Grundzüge der Prosasyntax* Stilprägende Entwicklungen vom Althochdeutschen zum Neuhochdeutschen (pp. 4–64). De Gruyter. (1st ed.).

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1995). *Dimensions of Register Variation. A Cross Linguistic Comparison*. Cambridge: Cambridge University Press.

Biber, D. (2006). *University Language: A Corpus-Based Study of Spoken and Written Registers* volume 23 of *Studies in Corpus Linguistics*. Amsterdam/Philadelphia: John Benjamins Publishing.

Biber, D., & Clark, V. (2002). Historical shifts in modification patterns with complex noun phrase structures. *English Historical Morphology. Current Issues in Linguistic Theory*, *11*, 43–66.

Biber, D., & Finegan, E. (1989). Drift and the evolution of English style: A history of three genres. *Language*, *65*, 487–517.

Biber, D., & Finegan, E. (1997). Diachronic relations among speech-based and written registers in English. In T. Nevalainen, & L. Kahlas-Tarkka (Eds.), *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen* (pp. 253–276). Helsinki: Société Néophilologique.

Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, *9*, 2–20.

Biber, D., & Gray, B. (2011a). Grammatical change in the noun phrase: The influence of written language use. *English Language and Linguistics*, *15*, 223–250. doi: 10.1017/S1360674311000025.

Biber, D., & Gray, B. (2011b). The historical shift of scientific academic prose in English towards less explicit styles of expression: Writing without verbs. In V. Bathia, P. Sánchez, & P. Pérez-Paredes (Eds.), *Researching Specialized Languages* (pp. 11–24). Amsterdam: John Benjamins.

Biber, D., & Gray, B. (2016). *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Studies in English Language. Cambridge, UK: Cambridge University Press.

Biber, D., Grieve, J., & Iberri-Shea, G. (2009). Noun phrase modification. In *One Language, Two Grammars?: Differences between British and American English* (pp. 182–193). Cambridge University Press.

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

Bizzoni, Y., Degaetano-Ortlieb, S., Fankhauser, P., & Teich, E. (2020). Linguistic variation and change in 250 years of English scientific writing: A data-driven approach. *Frontiers in Artificial Intelligence, section Language and Computation*, . doi: https://doi.org/10.3389/frai.2020.00073.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bock, J. K. (1986). Meaning, sound, and syntax: Lexical priming in sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, (pp. 575–586).

Bock, J. K., & Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, *21*, 47–67. doi: 10.1016/0010-0277(85)90023-X.

Bock, K. J., & Irwin, D. E. (1980). Syntactic effects of information availability in sentence production. *Journal of Verbal Learning and Verbal Behavior*, *19*, 467–484. doi: 10.1016/S0022-5371(80)90321-7.

Bräuer, R. (2001). Sprachhistorische Periodisierungskriterien und ihre Anwendung in der deutschen Sprachgeschichte. In T. Roelcke (Ed.), *Die zeitliche Gliederung der deutschen Sprachgeschichte. (Dokumentation germanistischer Forschung, Bd. 4)* (pp. 219–232).

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In *Cognitive Foundations of Interpretation* (pp. 69–94). KNAW.

Brinton, L. J., & Arnovick, L. K. (2006). *The English Language: A Linguistic History*. Don Mills, Ont.: Oxford University Press.

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

Brush, S. G., Osler, M. J., & Spencer, J. B. (2019). Scientific Revolution. URL: `https://www.britannica.com/science/Scientific-Revolution`.

Chen, B., Ning, A., Bi, H., & Dunlap, S. (2008). Chinese subject-relative clauses are more difficult to process than the object-relative clauses. *Acta Psychologica*, *129*, 61–65.

Chen, X., Alexopoulou, T., & Tsimpli, I. (2021). Automatic extraction of subordinate clauses and its application in second language acquisition research. *Behavior Research Methods*, *53*, 803–817. doi: 10.3758/s13428-020-01456-7.

Chi, Z. (1999). Statistical properties of probabilistic context-free grammars. *Computational Linguistics*, *25*, 131–160.

Chovanec, J. (2012). Grammar in the law. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 1–8). Oxford, UK: Blackwell Publishing Ltd. doi: 10.1002/9781405198431.wbeal0482.

Crystal, D. (1969). *Prosodic Systems and Intonation in English*. Cambridge: Cambridge University Press.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1983). A language-specific comprehension strategy. *Nature*, *304*, 159–160. doi: 10.1038/304159a0.

Dal, I., & Eroms, H.-W. (2014). *Kurze deutsche Syntax auf historischer Grundlage*. Sammlung kurzer Grammatiken germanischer Dialekte. B: Ergänzungsreihe (4th ed.). Berlin: De Gruyter.

de Courson, B., & Baumard, N. (2019). Quantifying the Scientific Revolution. SocArXiv.

De Smet, H. (2006). A corpus of late modern English texts. *ICAME Journal*, *29*, 69–82.

Dear, P. (1985). Totius in verba: Rhetoric and authority in the early Royal Society. *Isis: An International Review Devoted to the History of Science and Its Cultural Influences*, *76*, 145–161.

Decker, R. E. (1974). *Patterns of Essay IV*. Boston: Little, Brown & Company.

Degaetano-Ortlieb, S., Kermes, H., Khamis, A., & Teich, E. (2019). An information-theoretic approach to modeling diachronic change in scientific English. In C. Suhr, T. Nevalainen, & I. Taavitsainen (Eds.), *From Data to Evidence in English Language Research* number 83 in Language and Computers (pp. 258–281). Brill. doi: 10.1163/9789004390652.

Degaetano-Ortlieb, S., & Teich, E. (2016). Information-based modeling of diachronic linguistic change: From typicality to productivity. In *Proceedings of the 10th LaTeCH Workshop at ACL* (pp. 165–173).

Degaetano-Ortlieb, S., & Teich, E. (2018). Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature at COLING2018* (pp. 22–33). Santa Fe, NM, USA.

Degaetano-Ortlieb, S., & Teich, E. (2019). Toward an optimal code for communication: The case of scientific English. *Corpus Linguistics and Linguistic Theory*, . doi: 10.1515/cllt-2018-0088.

DeGraff, M. (2001). On the origin of creoles: A Cartesian critique of neo-Darwinian linguistics. *Linguistic Typology*, *5*, 213–310.

Dekeyser, X. (1984). Relativizers in early Modern English: A dynamic quantitative study. *Historical Syntax*, *23*, 61.

Dekeyser, X. (1986). English contact clauses revisited: A diachronic approach. *Folia Linguistica Historica*, *20*, 107–120. doi: 10.1515/flih.1986.7.1.107.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*, 193–210. doi: 10.1016/j.cognition.2008.07.008.

Diller, H.-J., De Smet, H., Tyrkkö, J., & Flach, S. (2011). A European database of descriptors of English electronic texts. *The European English Messenger*, *19*, 21–35.

Drozd, L., & Seibicke, W. (1973). *Deutsche Fach- und Wissenschaftssprache : Bestandsaufnahme, Theorie, Geschichte*. (1st ed.). Wiesbaden: Brandstetter.

Du Bois, J. W. (1985). Competing motivations. *Iconicity in Syntax*, *6*, 343–365.

Ebert, R. P. (1986). *Historische Syntax des Deutschen* volume 2: 1300–1750 of *Germanistische Lehrbuchsammlung; 6*. Bern [u.a.]: Lang.

Ebert, R. P., Reichmann, O., Solms, H.-J., & Wegera, K.-P. (Eds.) (1993). *Frühneuhochdeutsche Grammatik*. Sammlung kurzer Grammatiken germanischer Dialekte – A. Tübingen: Niemeyer.

Edmonds, B. (1995). What is complexity? – The philosophy of complexity per se with application to some examples in evolution. In *The Evolution of Complexity*. Kluwer, Dordrecht.

Eggers, H. (1977). *Deutsche Sprachgeschichte* volume 4: Das Neuhochdeutsche of *Rowohlts Deutsche Enzyklopädie ; 375*. Reinbek bei Hamburg: Rowohlt.

Eggers, H. (1986). *Deutsche Sprachgeschichte* volume 2: Das Frühneuhochdeutsche und das Neuhochdeutsche of *Rowohlts Deutsche Enzyklopädie ; 426*. Reinbek bei Hamburg: Rowohlt.

Erben, J. (1984). *Deutsche Syntax: Eine Einführung* volume 12 of *Germanistische Lehrbuchsammlung*. Peter Lang GmbH, Internationaler Verlag der Wissenschaften.

Ernst, P. (2021). *Deutsche Sprachgeschichte: Eine Einführung in die diachrone Sprachwissenschaft des Deutschen*. UTB Basics (3rd ed.). Wien: Facultas.

Fedorenko, E., Woodbury, R., & Gibson, E. (2013). Direct evidence of memory retrieval as a source of difficulty in non-local dependencies in language. *Cognitive Science*, *37*, 378–394. doi: 10.1111/cogs.12021.

Ferrer i Cancho, R. (2004). Euclidean distance between syntactically linked words. *Physical Review E*, *70*, 056135. doi: 10.1103/PhysRevE.70.056135.

Fischer, S., Knappen, J., Menzel, K., & Teich, E. (2020). The Royal Society Corpus 6.0: Providing 300+ years of scientific writing for humanistic study. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 794–802). Marseille, France: European Language Resources Association.

Fleischer, J. (2002). Die Syntax von Pronominaladverbien in den Dialekten des Deutschen. *Zeitschrift für Dialektologie und Linguistik: Beihefte*, (p. 1).

Fleischer, J. (2013). Relativsätze in den Dialekten des Deutschen: Vergleich und Typologie. *Linguistik Online*, *24*. doi: 10.13092/lo.24.642.

Frank, S. L. (2013). Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, *5*, 475–494. doi: 10.1111/tops.12025.

Frank, S. L., & Frank, S. L. (2009). Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *31*, 7.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2013). Word surprisal predicts N400 amplitude during reading. (p. 123). Sofia, Bulgaria: Association for Computational Linguistics.

Futrell, R., Levy, R. P., & Gibson, E. (2020). Dependency locality as an explanatory principle for word order. *Language*, *96*, 371–412. doi: 10.1353/lan.2020.0024.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*, 10336–10341. doi: 10.1073/pnas.1502134112.

Fyfe, A., Squazzoni, F., Torny, D., & Dondio, P. (2020). Managing the growth of peer review at the Royal Society journals, 1865–1965. *Science, Technology, & Human Values*, *45*, 405–429. doi: 10.1177/0162243919862868.

Gell-Mann, M. (1992). Complexity and complex adaptive systems. In J. A. Hawkins, & M. Gell-Mann (Eds.), *The Evolution of Human Languages*. Reading: Addison-Wesley.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 199–206).

Geyken, A., Boenig, M., Haaf, S., Jurish, B., Thomas, C., & Wiegand, F. (2018). 10. Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN. In H. Lobin, R. Schneider, & A. Witt (Eds.), *Digitale Infrastrukturen für die germanistische Forschung* (pp. 219–248). De Gruyter. doi: 10.1515/9783110538663-011.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*, 1–76.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, Language, Brain*, *2000*, 95–126.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*, 389–407. doi: 10.1016/j.tics.2019.02.003.

Gibson, E., & Wu, H.-H. I. (2013). Processing Chinese relative clauses in context. *Language and Cognitive Processes*, *28*, 125–155. doi: 10.1080/01690965.2010.536656.

Gildea, D., & Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science*, *34*, 286–310. doi: 10.1111/j.1551-6709.2009.01073.x.

Giunta, C. (2023). Glossary of Archaic Chemical Terms. URL: `https://web.lemoyne.edu/~giunta/archema.html`.

Givón, T. (2017). *The Story of Zero*. John Benjamins Publishing Company.

Görlach, M. (2001). *Eighteenth-Century English*. Sprachwissenschaftliche Studienbücher. Heidelberg: Winter.

Gotti, M. (2003). *Specialized Discourse: Linguistic Features and Changing Conventions*. Bern: P. Lang.

Gottsched, J. (1748). *Chr. Grundlegung einer deutschen Sprachkunst*.

Grafmiller, J., & Shih, S. (2011). New approaches to end weight. *Variation and Typology: New Trends in Syntactic Research*, *26*.

Gray, B. (2015). On the complexity of academic writing: Disciplinary variation and structural complexity. In *Corpus-based Research in Applied Linguistics* (pp. 49–78). John Benjamins.

Grice, H. P. (1975). Logic and conversation. *Speech Acts*, *3*, 45–58.

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, *29*, 261–290.

Guy, G. R., & Bayley, R. (1995). On the choice of relative pronouns in English. *American Speech*, *70*, 148–162. doi: 10.2307/455813.

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

Haiman, J. (1983). Iconic and economic motivation. *Language*, *59*, 781–819.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies* (pp. 1–8). Association for Computational Linguistics.

Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, *32*, 101–123.

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*, 643–672. doi: 10.1207/s15516709cog0000_64.

Hall, A. R. (1954). *The Scientific Revolution, 1500–1800: the Formation of the Modern Scientific Attitude*. Longmans.

Halliday, M. (1988). On the language of physical science. In M. Ghadessy (Ed.), *Registers of Written English: Situational Factors and Linguistic Features* (pp. 162–177). London: Pinter.

Halliday, M., & Martin, J. R. (1993). *Writing Science: Literacy and Discursive Power*. London: Falmer Press.

Halliday, M., & Ruqaiya Hasan (1985). *Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective*. Oxford: Oxford University Press.

Halliday, M. A. K. (1985). *Spoken and Written Language*. Deakin University.

Halliday, M. A. K., & Webster, J. (2004). *The Language of Science* volume 5 of *Collected Works of M.A.K. Halliday*. London; New York: Continuum.

Hartweg, F., & Wegera, K.-P. (2005). *Frühneuhochdeutsch: Eine Einführung in die deutsche Sprache des Spätmittelalters und der frühen Neuzeit*. Germanistische Arbeitshefte 33 (2nd ed.). Tübingen: Max Niemeyer Verlag.

Haspelmath, M. (1999). Optimality and diachronic adaptation. *Zeitschrift für Sprachwissenschaft*, *18*, 180–205. doi: 10.1515/zfsw.1999.18.2.180.

Haspelmath, M. (2021). Explaining grammatical coding asymmetries: Form–frequency correspondences and predictability. *Journal of Linguistics*, *57*, 605–633. doi: 10.1017/S0022226720000535.

Hawkins, J. A. (1994). *A Performance Theory of Order and Constituency*. Cambridge University Press.

Hawkins, J. A. (2003). Efficiency and complexity in grammars: Three general principles. *The Nature of Explanation in Linguistic Theory*, *121*, 121–152.

Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. Oxford University Press.

Hawkins, J. A. (2009). An efficiency theory of complexity and related phenomena. In *Language Complexity as an Evolving Variable*. Oxford University Press volume 13.

Hawkins, J. A. (2014). *Cross-Linguistic Variation and Efficiency*. Oxford: OUP.

Helbig, G., & Buscha, J. (2001). *Deutsche Grammatik: Ein Handbuch für den Ausländerunterricht*. Langenscheidt.

Henderson, J. M., Choi, W., Lowder, M. W., & Ferreira, F. (2016). Language structure in the brain: A fixation-related fMRI study of syntactic surprisal in reading. *Neuroimage*, *132*, 293–300.

Heringer, H. J., Strecker, B., & Wimmer, R. (1980). *Syntax: Fragen, Lösungen, Alternativen*. Fink.

Hinrichs, E., Feldweg, H., Boyle-Hinrichs, M., & Hauser, R. (1995). *Abschlüssbericht ELWIS. Korpusunterstützte Entwicklung lexikalischer Wissensbasen für die Computerlinguistik*. Technical Report Universität Tübingen Tübingen.

Hofmeister, P., Jaeger, T. F., Sag, I. A., Arnon, I., & Snider, N. (2007). Locality and accessibility in wh-questions. *Roots: Linguistics in Search of its Evidential Base*, (pp. 185–206).

Hofmeister, P., & Sag, I. A. (2010). Cognitive constraints and island effects. *Language*, *86*, 366–415.

Honnibal, M., & Ines, M. (2022). SpaCy. URL: `https://spacy.io/models/de`.

Horn, L. (1984). Towards a new taxonomy for pragmatic inference: Q-based and R-based implicature. *Meaning, Form and Use in Context*, (p. 123).

Hsiao, F., & Gibson, E. (2003). Processing relative clauses in Chinese. *Cognition*, *90*, 3–27.

Huber, M. (2017). Structural and sociolinguistic factors conditioning the choice of relativizers in Late Modern English: A diachronic study based on the Old Bailey Corpus. *Nordic Journal of English Studies*, *16*, 74–119.

Hudson, R. (1995). Measuring syntactic difficulty. Manuscript, University College, London.

Hudson, R. A. (1991). *English Word Grammar*. B. Blackwell.

Hundt, M., Denison, D., & Schneider, G. (2012). Relative complexity in scientific discourse. *English Language and Linguistics*, *16*, 209–240. doi: 10.1017/S1360674312000032.

Ingels, M. (1985). Socio-historical aspects of relativization in 16th century English. Unpublished dissertation. University of Leuven.

Jaeger, T. F., & Tily, H. (2011). On language 'utility': Processing complexity and communicative efficiency. *WIREs Cognitive Science*, *2*, 323–335. doi: 10.1002/wcs.126.

Johansson, C. (2006). Relativizers in 19th-century English. In M. Kytö, M. Rydén, & E. Smitterberg (Eds.), *Nineteenth-century English: Stability and Change* (pp. 136–182). Cambridge: Cambridge University Press.

Johansson, C. (2012). Relativization in Early Modern English: Written versus speech-related genres. In *English Historical Linguistics: An International Handbook* (pp. 776–790). volume 1.

Johansson, C., & Geisler, C. (1998). Pied piping in spoken English. *Language and Computers*, *23*, 67–82.

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, *5*, 206–213. doi: 10.1080/09296179808590128.

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

Jurish, B. (2012). CAB Web Service. URL: `https://www.deutschestextarchiv.de/demo/cab/`.

Jurish, B. (2020). DTA Tokwrap. URL: `https://kaskade.dwds.de/~moocow/software/dta-tokwrap/`.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*, 329.

Juzek, T. S., Fischer, S., Krielke, P., Degaetano-Ortlieb, S., & Teich, E. (2019a). Challenges of parsing a historical corpus of Scientific English. In *Historical Corpora and Variation (HiCoV)*.

Juzek, T. S., Fisher, S., Krielke, P., Degaetano-Ortlieb, S., & Teich, E. (2019b). Annotation quality assessment and error correction in diachronic corpora: Combining pattern-based and machine learning approaches. In *Societas Linguistica Europea* (p. 668).

Juzek, T. S., Krielke, M.-P., & 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)* (pp. 109–119).

Keenan, E. L., & Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, *8*, 63–99.

Keenan, E. L., & Hawkins, S. (1987). The psychological validity of the accessibility hierarchy. *Universal Grammar*, *15*, 60–85.

Keller, R. et al. (1994). *On Language Change: The Invisible Hand in Language*. London: Routledge.

Keller, R. E. (1978). *The German Language*. London/Boston: Humanities Press International.

Kelly, M. H., Bock, J. K., & Keil, F. C. (1986). Prototypicality in a linguistic context: Effects on sentence structure. *Journal of Memory and Language*, (p. 25:59–74).

Kintsch, W. (1974). *The Representation of Meaning in Memory*. Lawrence Erlbaum.

Knappen, J. (2022). Gute Sätze Tools. URL: `https://github.com/SFB1102/B1-gute-saetze`.

Konopka, M. (1996). *Strittige Erscheinungen der deutschen Syntax im 18. Jahrhundert*. Number 173 in Reihe Germanistische Linguistik. Tübingen: Niemeyer.

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*, 91–120. doi: 10.15845/bells.v11i1.3440.

Krielke, M.-P., Talamo, L., Fawzi, M., & Knappen, J. (2022). Tracing syntactic change in the scientific genre: Two universal dependency-parsed diachronic corpora of scientific English and German. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 4808–4816).

Kučera, H., & Francis, W. N. (1967). *Computational Analysis of Present-Day American English*. Providence, RI, USA: Brown University Press.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*, 79–86.

Langacker, R. W. (1977). *Syntactic Reanalysis: Mechanism of Syntactic Change*. University of Texas Press, Austin.

Lau, E., & Tanaka, N. (2021). The subject advantage in relative clauses: A review. *Glossa: A Journal of General Linguistics*, *6*. doi: 10.5334/gjgl.1343.

Leech, G., Mair, C., Smith, N., & Hundt, M. (2009). *Changes in Contemporary English: A Corpus-Based Study*. Cambridge University Press.

Lehmann, C. (1984). *Der Relativsatz: Typologie seiner Strukturen, Theorie seiner Funktionen, Kompendium seiner Grammatik*. Number 3 in Language Universals Series. Tübingen: G. Narr.

Levey, S. (2006). Visiting London relatives. *English World-Wide*, *27*, 45–70.

Levinson, S. C. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. MIT Press.

Levshina, N. (2018). *Towards a Theory of Communicative Efficiency in Human Languages*. Ph.D. thesis Universität Leipzig. doi: 10.5281/zenodo.1542857.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177.

Levy, R. P., & Keller, F. (2013). Expectation and locality effects in German verb-final structures. *Journal of Memory and Language*, *68*, 199–222. doi: 10.1016/j.jml.2012.02.005.

Lin, Y., & Garnsey, S. M. (2011). Animacy and the resolution of temporary ambiguity in relative clause comprehension in Mandarin. In H. Yamashita, Y. Hirose, & J. L. Packard (Eds.), *Processing and Producing Head-final Structures* Studies in Theoretical Psycholinguistics (pp. 241–275). Dordrecht: Springer Netherlands. doi: 10.1007/978-90-481-9213-7_12.

Lindelöf, A. (1997). Prepositional relatives in English: A diachronic study of which and who(m) in early Modern English texts. Unpublished paper. Department of English, Göteborg University.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, *9*, 159–191. doi: 10.17791/JCS.2008.9.2.159.

Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, *21*, 171–193. doi: 10.1016/j.plrev.2017.03.002.

Love, T. E., & Swinney, D. A. (1998). The influence of canonical word order on structural processing. In *Sentence Processing: A Crosslinguistic Perspective* (pp. 153–166). Brill.

Ludiková, M. (1987). On the semantics of pronominal adverbs from the quantitative aspect. *Prague Studies in Mathematical Linguistics*, *9*, 53–64.

Lui, M., & Baldwin, T. (2012). Langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations* (pp. 25–30). Jeju Island, Korea: Association for Computational Linguistics.

Maas, U. (2012). *Bildungsverhältnisse im 16. Jahrhundert*. Brill Fink. doi: 10.30965/9783846752722_017.

Maclean, I. (1963). Strategien der Kommunikation von Naturwissen und Medizin: Zeitschriften gelehrter Akademien in der frühen Neuzeit. (pp. 37–67). Stuttgart: Wissenschaftliche Verlagsgesellschaft volume 81 of *Acta Historica Leopoldina*. (1st ed.). doi: 10.26164/leopoldina_10_00710.

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*, 255–308. doi: 10.1162/coli_a_00402.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. URL: `http://mallet.cs.umass.edu`.

Mellinkoff, D. (2004). *The Language of the Law*. Wipf and Stock.

Menzel, K., Knappen, J., & Teich, E. (2021). Generating linguistically relevant metadata for the Royal Society Corpus. *Research in Corpus Linguistics*, (p. 18).

Milin, P., Kuperman, V., Kostic, A., & Baayen, R. H. (2009). Paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation. *Analogy in Grammar: Form and Acquisition*, (pp. 214–252).

Moscoso del Prado Martín, F., Kostic, A., & Baayen, R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*, (pp. 1–18). doi: 10.1016/j.cognition.2003.10.015.

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.

Mufwene, S. S., Coupé, C., & Pellegrino, F. (2017). *Complexity in Language: Developmental and Evolutionary Perspectives*. Cambridge University Press.

Müller, G. (2000). Das Pronominaladverb als Reparaturphänomen. *Linguistische Berichte*, (pp. 139–178).

Mustanoja, T. F. (1960). Middle English Syntax, Part I: Parts of Speech. *Mémoires de la Soclété Néophilologique de Helsinki*, *23*.

Nedoluzhko, A., & Lapshinova-Koltunski, E. (2018). Pronominal adverbs in German and their equivalents in English, Czech and Russian: Evidence from the Parallel

Corpus. In *Computer Linguistics and Intellectual Technologies: Materials of the Annual International Conference "Dialogue", Moscow* (p. 2018). volume 17.

Negele, M. (2012). *Varianten der Pronominaladverbien im Neuhochdeutschen*. De Gruyter.

Nerius, D. (1967). *Untersuchungen zur Herausbildung einer nationalen Norm der deutschen Literatursprache im 18. Jahrhundert*. Halle.

Nevalainen, T., & Raumolin-Brunberg, H. (2002). The rise of the relative who in Early Modern English. *Relativisation on the North Sea Littoral*, (pp. 109–121).

Nevalainen, T., & 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* (pp. 171–216). De Gruyter Mouton.

Newmeyer, F. J., & Preston, L. B. (2014). *Measuring Grammatical Complexity*. Oxford University Press, USA.

Nicenboim, B., Vasishth, S., Gattei, C., Sigman, M., & Kliegl, R. (2015). Working memory differences in long-distance dependency resolution. *Frontiers in Psychology*, *6*, 312. doi: 10.3389/fpsyg.2015.00312.

Nichols, J. (1992). *Linguistic Diversity in Space and Time*. University of Chicago Press.

Onishi, K. H., Murphy, G. L., & Bock, K. (2008). Prototypicality in sentence production. *Cognitive Psychology*, *56*, 103–141. doi: 10.1016/j.cogpsych.2007.04.001.

Osminkin, A. (2020). Pronominal adverbs based on here-, there-, and where- as textual connectors in legal discourse. *International Journal of Law, Language & Discourse*, *8*, 57–85.

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. *Language Policy*, *19*, 235–258. doi: 10.1007/s10993-019-09530-3.

Piirainen, I. T. (1980). *Deutsche Standardsprache des 17./18. Jahrhunderts*. Max Niemeyer Verlag.

Pittner, K. (2008). Schlecht dran oder gut drauf? – Überlegungen zur Grammatikalisierung und Akzentuierung von Pronominaladverbien. *Deutsche Sprache*, (pp. 74–94). doi: 10.37307/j.1868-775X.2008.01.06.

von Polenz, P. (1978). *Geschichte der deutschen Sprache*. Berlin/New York: Sammlung Göschen.

von Polenz, P. (1991). *Einführung, Grundbegriffe, Deutsch in der Frühbürgerlichen Zeit*. De Gruyter. doi: 10.1515/9783110853421.

von Polenz, P. (1999). *Deutsche Sprachgeschichte vom Spätmittelalter bis zur Gegenwart, Bd III, 19. und 20. Jahrhundert*. De Gruyter.

Pörksen, U. (1986). *Deutsche Naturwissenschaftssprachen*. Tübingen: Narr.

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

Rayner, K. (1977). Visual attention in reading: Eye movements reflect cognitive processes. *Memory & Cognition*, *5*, 443–448.

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, *14*, 191–201. doi: 10.3758/BF03197692.

Reali, F., & Christiansen, M. H. (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language*, *57*, 1–23. doi: 10.1016/j.jml.2006.08.014.

Rijkhoff, J. (1990). Explaining word order in the noun phrase. *Linguistics*, *28*, 5–42.

Rissanen, M. (1984). The choice of relative pronouns in 17th century American English. *Historical Syntax*, *23*, 417.

Roelcke, T. D. (2020). *Fachsprachen*. Grundlagen der Germanistik; Basics (4th ed.). Berlin: Erich Schmidt Verlag.

Rohdenburg, G. (1996). Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics*, *7*, 149–182.

Rohdenburg, G. (2006). Evolution of the English complementation system. *Syntax, Style and Grammatical Norms: English from 1500–2000*, *39*, 143.

Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, *57*, 348–379. doi: 10.1016/j.jml.2007.03.002.

Romaine, S. (1980). The relative clause marker in Scots English: Diffusion, complexity, and style as dimensions of syntactic change. *Language in Society*, *9*, 221–247.

Romaine, S. (1982). *Socio-Historical Linguistics: Its Status and Methodology*. Cambridge Studies in Linguistics, 34. Cambridge: Cambridge University Press.

Romaine, S. (1998). *The Cambridge History of the English Language. Volume 4, 1776–1997*. Cambridge: University Press.

Rydén, M. (1966). *Relative Constructions in Early Sixteenth Century English: With Special Reference to Sir Thomas Elyot*. Acta Universitatis Upsaliensis. Uppsala: Almquist & Wiksells.

Santorini, B. (1990). *Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision)*. Technical Report University of Pennsylvania (CIS).

Sapir, E. (1921). An introduction to the study of speech. *Language*, *1*.

Scaglione, A. D. (1981). *Komponierte Prosa von der Antike bis zur Gegenwart*. Klett-Cotta.

Scherer, W. (1875). *Geschichte der deutschen Dichtung im elften und zwölften Jahrhundert*. Quellen und Forschungen zur Sprach- und Kulturgeschichte der germanischen Völker; 12. Strassburg: Trübner.

Schildt, J. (1984). *Abriss der Geschichte der deutschen Sprache: zum Verhältnis von Gesellschafts- und Sprachgeschichte*. Akademie-Verlag.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing* (pp. 44–49). Manchester, UK.

Schottel, J. G. (1967). *Ausführliche Arbeit von Der Teutschen Haubt Sprache. 1663* volume 11. Niemeyer.

Sekerina, I. A. (2003). Scrambling and processing: Dependencies, complexity, and constraints. In S. Karimi (Ed.), *Word Order and Scrambling* (pp. 301–324). Wiley-Blackwell volume 4.

Semenjuk, N. N. (1972). Zustand und Evolution der grammatischen Normen des Deutschen in der 1. Hälfte des 18. Jahrhunderts. *Studien zur Geschichte der deutschen Sprache*, *49*, 79–166.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Shapin, S. (2018). *The Scientific Revolution*. University of Chicago Press.

Sinclair, J. H. (Ed.) (1995). *Collins COBUILD English Dictionary*. The COBUILD Series (completely new ed.). London: Harper Collins.

Sladen, C. F. (1917). *The Approach of Academic to Spoken Style in German: A Study in Popular Scientific Prose from 1850 to 1914*. Ph.D. thesis University of Pennsylvania Philadelphia.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*, 302–319. doi: https://doi.org/10.1016/j.cognition.2013.02.013.

Staples, S., Egbert, J., Biber, D., & Gray, B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication*, *33*, 149–183.

Steels, L., & Beuls, K. (2017). How to explain the origins of complexity in language: A case study for agreement systems. In S. Mufwene, C. Coupé, & F. Pellegrino (Eds.), *Complexity in Language: Developmental and Evolutionary Perspectives* (pp. 30–47). Cambridge: Cambridge University Press.

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 197–207).

Strang, B. (1970). *A History of the English Language*. London: Routledge.

Suárez-Gómez, C. (2008). Strategies in competition: Demonstratives and interrogatives as relativizers in the history of English. *English Studies*, *89*, 339–350.

Suárez-Gómez, C. (2012). Clause linkage across time and genres in early English: A preliminary approach to relative clauses. *Studia Neophilologica*, *84*, 138–150.

Tagliamonte, S. (2002). Variation and change in the British relative marker system. *Relativisation on the North Sea Littoral*, (pp. 147–165).

Tagliamonte, S., Smith, J., & Lawrence, H. (2005). No taming the vernacular! Insights from the relatives in northern Britain. *Language Variation and Change*, *17*, 75–112. doi: 10.1017/S0954394505050040.

Teich, E., Fankhauser, P., Degaetano-Ortlieb, S., & Bizzoni, Y. (2021). Less is more/more diverse: On the communicative utility of linguistic conventionalization. *Frontiers in Communication*, *5*, 142. doi: 10.3389/fcomm.2020.620275.

Teich, M. (2015). *The Scientific Revolution Revisited*. Open Book Publishers. doi: 10.11647/OBP.0054.

The Royal Society (2022). History of the Royal Society. URL: `https://royalsociety.org/about-us/history/#:~:text=The%20Royal%20Society's%20motto%20'Nullius,to%20facts%20determined%20by%20experiment`.

The Royal Society of Chemistry (2023). Hydrogen. URL: `https://www.rsc.org/periodic-table/element/1/hydrogen#`.

Tiersma, P. M. (1999). *Legal Language*. University of Chicago Press.

Tottie, G. (1995). The man Ø I love: An analysis of factors favouring zero relatives in written British and American English. *Stockholm Studies in English*, *85*, 201–215.

Tottie, G., & Harvie, D. (2000). It's all relative: Relativization strategies in early African American English. *Language in Society – Oxford*, *28*, 198–232.

Trotta, J. (2000). *Wh-Clauses in English: Aspects of Theory and Description*. 34. Rodopi.

Tschirch, F. (1989). *Geschichte der deutschen Sprache* volume 2: Entwicklung und Wandlungen der deutschen Sprachgestalt vom Hochmittelalter bis zur Gegenwart of *Grundlagen der Germanistik ; 9*. (3rd ed.). Berlin: Schmidt.

Universal Dependencies.org (2023a). Fixed Relation. URL: `https://universaldependencies.org/de/dep/fixed.html`.

Universal Dependencies.org (2023b). Universal Dependencies Tools. URL: `https://universaldependencies.org/tools.html`.

Utah State University (2020). Aquatic Macroinvertebrates. URL: `https://extension.usu.edu/waterquality/learnaboutsurfacewater/propertiesofwater/aquaticmacros`.

Valle, E. (1997). A scientific community and its texts: A historical discourse study. In *The Construction of Professional Discourse* (pp. 76–98). Essex: Longman.

Van den Eynden, N. (1996). *Aspects of Preposition Placement in English*. Frankfurt: Lang.

Wasow, T. (2002). *Postverbal Behavior*. CSLI Stanford, CA.

Weaver, C. A., & Kintsch, W. (1991). Expository text. In *Handbook of Reading Research*. New York: Longman volume 2.

Wiese, B. (2008). Kasusdifferenzierung in der neuhochdeutschen Nominalgruppe. In *Arbeitspapier Grammatik des Deutschen im europäischen Vergleich. Mannheim*.

Williams, C. (2007). *Tradition and Change in Legal English: Verbal Constructions in Prescriptive Texts* volume 20. Peter Lang.

Windsor, F. M., Tilley, R. M., Tyler, C. R., & Ormerod, S. J. (2019). Microplastic ingestion by riverine macroinvertebrates. *Science of the Total Environment*, *646*, 68–74. doi: 10.1016/j.scitotenv.2018.07.271.

Wolff, G. (1991). Deutsche Sprachgeschichte. *Informationen Deutsch als Fremdsprache*, *18*, 696–702.

Wydick, R. C., & Sloan, A. E. (2005). *Plain English for Lawyers* volume 4. Durham, NC: Carolina Academic Press.

Yáñez-Bouza, N. (2011). ARCHER past and present (1990–2010). *ICAME Journal*, *35*, 205–236.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Eoclogy*. Cambridge: Addison-Wesley Press.