The Interplay of Noun Phrase Complexity and Modification Type in Scientific Writing

Isabell Landwehr
Department of Language Science and Technology
Saarland University
isabell.landwehr@uni-saarland.de

Abstract

We investigate the interplay of noun phrase (NP) complexity and modification type, namely the choice between pre- and postmodification, using a corpus-based approach. Our dataset is the Royal Society Corpus (RSC; Fischer et al., 2020), a diachronic corpus of English scientific writing. We find that the number of dependents, length of the head noun and distance to the head noun's own syntactic head (typically the main verb) affect the likelihood of pre- vs. postmodification: NPs with more dependents are more likely to be premodified, NPs with a longer head noun and a head noun closer to its own head are more likely to be postmodified. In addition, we find an effect of syntactic role and definiteness as well as time: The likelihood of premodification over postmodification increases with time and subject NPs as well as indefinite NPs are more likely to be premodified than NPs in other syntactic roles or definite NPs.

1 Introduction

Language use has been argued to be shaped by optimization constraints (e.g. Levshina, 2022), such as minimizing dependency length between syntactic heads and dependents (e.g. Gibson, 1998; Gibson et al., 2000). This has also been posited for the register of English scientific writing (Degaetano-Ortlieb and Teich, 2022), in which complex noun phrases (e.g. NPs consisting of a head noun modified by several dependents) are a key feature (Halliday, 1988). English allows both premodification (e.g., in the form of nouns or adjectives, see Example 1) and post-modification (e.g. prepositional phrases or finite and non-finite relative clauses, see Example 2) in the noun phrase. The two types of modification may also occur at the same time (Example 3). [1]

(1) However, in this case we may proceed to calculate **the total plasma velocity** directly [...]. *(RSC, rsta_1996_0136)*

(2) But when **velocity relative to aether** was finally abandoned [...]. *(RSC, rsbm_1942_0016)*

(3) So far we have calculated **the flow velocity normal to the field lines** [...]. *(RSC, rsta_1996_0136)*

Previous studies on scientific writing have observed a diachronic shift from postmodification to premodification of the noun phrase (Degaetano-Ortlieb, 2021). Premodification results in more compressed structures than postmodification, which is particularly the case for nouns premodified by other nouns, i.e. compounds: Not only does a compound like *the plasma velocity* contain fewer words compared to a prepositional phrase (*the velocity of plasma*) or a relative clause (*the velocity which plasma possesses*), it also makes the semantic relationship between modifier and head implicit: The relationship between *plasma* and *velocity* could theoretically be interpreted as *plasma* ***has*** *velocity* (similarly to *eye color*), *velocity* ***consists of*** *plasma* (similarly to *stone pillar*) or *velocity* ***found in*** *plasma* (similarly to *forest animal*). [2] Selecting the correct semantic relation out of several competing ones is a crucial task in compound processing and high entropy of possible relations increases processing difficulty (Benjamin and Schmidtke, 2023). Moreover, the internal embedding structure of compounds may be ambiguous as well. A three-constituent compound such as *energy*

[1] Examples are taken from the Royal Society Corpus (RSC; Fischer et al., 2020).

[2] A detailed discussion and annotation scheme of semantic relations between constituents can be found in Gagné and Shoben (1997) or Ó Séaghdha (2007), among others.

Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025), pages 72–82

flow velocity could refer to the *velocity of the energy flow* or the *flow velocity of energy*. This means that, while premodification streamlines linguistic structures, it also adds a new level of complexity. In addition, the choice of modification also influences other features of linguistic complexity like dependency length (see Section 2.2).

The present study aims to investigate how the increased complexity introduced by highly compressed structures interacts with other aspects of NP complexity. Premodification has become very common in scientific writing, but does this hold equally for all types of NPs, regardless of their complexity (e.g. in terms of number of dependents)? We assume that language users, in general but particularly in scientific writing, aim to maintain communicative efficiency (Levshina, 2022; Degaetano-Ortlieb and Teich, 2022), for instance by avoiding excessive complexity. Given processing constraints, we investigate the hypothesis that more complex NPs (e.g. with larger numbers of dependents) tend to be postmodified rather than premodified. Taking a corpus-based approach, we utilize Universal Dependencies annotation (de Marneffe et al., 2021) to consider different dependency-based complexity features. In our statistical analysis, we find that several complexity features influence modification type: In contrast to our original hypothesis, we find that a higher dependency number is associated with a greater likelihood of premodification. Features like larger distance to the verbal head and greater head noun length, on the other hand, are all associated with a greater likelihood of postmodification. Discourse status and syntactic role affect modification type as well, with indefinite NPs and subject NPs being more likely to be premodified. We also observe an effect of time in line with previous studies, namely an increased likelihood of premodification in later years.

This paper is structured as follows: Section 2 introduces relevant previous work on scientific writing and linguistic complexity, taking both psycholinguistic and theoretical approaches into account. It also motivates the selection of complexity features included in the analysis, while Section 3 describes our dataset and the preprocessing steps. Section 4 presents the statistical analysis, with a discussion in subsection 4.4. Section 5 on limitations and possible future research wraps up the paper.

2 Background and Rationale

2.1 Complex Noun Phrases in Scientific Writing

We conceptualize the general writing process in a similar way as described by Flower and Hayes (1981) and Hayes and Flower (1987): A writer first plans what they are going to write about (e.g. about the concept of *plasma velocity*). They then translate their ideas into syntactic structure, generating a sentence. During this step, NPs are used to encode the main concepts (e.g. *velocity*), possible modifiers further elaborate these core concepts (e.g. *plasma* and *total*). If modifiers are included, a choice between premodification, postmodification or a combination of both needs to be made at this stage. In a final step, the writer revises and edits the produced text. These three main steps can overlap and be repeated recursively. In scientific writing, complex NPs fulfill a central role for encoding concepts: Nominalization is a key feature of this register and NPs frequently describe very technical and specialized concepts (Halliday, 1988; Banks, 2008). Historically, there has been a development from an emphasis on clausal structures to an emphasis on phrasal structures, allowing information to be conveyed in a more compressed way (Biber and Gray, 2011). This is particularly exemplified by the case of compounds, which are informationally denser than their prepositional counterparts, and which have increased in frequency over time (Degaetano-Ortlieb, 2021). In this way, scientific English writing evolved to be optimized for written communication among experts (Degaetano-Ortlieb and Teich, 2022).

Writing a formal text such as a scientific article is also an audience-directed process (Hayes and Flower, 1987): The writer aims to accommodate the needs of the potential reader(s) and to make the text understandable. This means that possible processing demands on the reader need to be considered as well.

2.2 Processing Complex Structures

Previous studies have analyzed linguistic complexity from different perspectives. The processing cost associated with complex structures and the influence of complexity features on constituent order have been of particular interest.

A frequently investigated feature of complexity is dependency length, which describes the distance between a syntactic head and its dependent(s).

Greater distance has been associated with increased processing cost: According to Dependency Locality Theory (DLT; Gibson, 1998; Gibson et al., 2000), greater distance means that the prediction about upcoming material needs to be kept in memory for a longer time. This increases the cost for maintaining the prediction and for finally integrating it into the mental syntactic representation. Dependency length has therefore been proposed to measure processing difficulty (Liu, 2008). Support for DLT comes from various studies: Gibson (1998) showed its ability to account for different complexity phenomena, such as the processing of subject- versus object-extracted relative clauses. Liu (2008) analyzed dependency distance in a corpus study covering 20 languages and found a trend towards minimization of average dependency distance. Demberg and Keller (2008) found that DLT successfully predicts the reading times for nouns, while Temperley (2007) tested its predictions from a production perspective. Accordingly, the principle of Dependency Length Minimization (DLM) has been proposed, which posits that language users aim to place syntactic heads and dependents in proximity to each other (Futrell, 2019) and is often regarded as a linguistic universal (Liu et al., 2017). Similarly, dependency locality has been associated to information locality, e.g. by Levshina (2022), who argues that language users aim to minimize dependency length in order to maintain communicative efficiency.

Another complexity feature is the length of syntactic constituents, which has been found to affect constituent order: Behaghel (1909) already observed that long, complex phrases tend to occur at the ends of clauses (called *end-weight* in other studies, see e.g. Eitelmann (2016)). Discourse status needs to be considered as well: *Given* information tends to precede *New* information (Gundel et al., 1988; Prince, 1992). Arnold et al. (2000) found that heavy and new NPs tend to be postponed in the sentence, giving the speaker more time to plan the utterance and easing memory load on the listener. Syntactic role has also been considered when investigating dependency length, constituent length and discourse status: Temperley (2007), for instance, found that in written English, direct objects tend to be longer than subjects, and that postmodifying adverbial clauses tend to be longer than premodifying adverbial clauses.

In addition, word length itself has also been shown to affect processing (Baddeley et al., 1975; Jalbert et al., 2011; Guitard et al., 2018): Shorter words are recalled better than longer words, indicating a higher load on working memory associated with longer words.

Focusing specifically on the effect of NP structure on language understanding, an experimental study by Mota and Igoa (2017) compared simple NPs, which consisted of a series of coordinate NPs, and complex NPs, which contained embedded prepositional phrases. They found that language comprehenders were sensitive to the NP complexity, but only in the case of subject NPs.

2.3 Rationale

We investigate how different features of NP complexity interact with modification type. We selected the features based on previous literature (see Sections 2.1 and 2.2) and chose to consider all of them in order to limit possible confounding effects and improve the validity of our results. Modification type may be influenced by the overall **dependency length**, the distance between the head noun of the NP to its own verbal head. If many tokens already intervene between an NP and its head, the choice between pre- and postmodification can further increase this distance, depending on the **syntactic role** of the NP: A subject NP's distance to the head is increased by postmodifiers, an object or oblique's distance to the head is increased by premodification. In Example 2, for instance, the distance between *velocity*, the subject NP's head noun and *abandoned*, its verbal head, is 6 steps. Without the subject NP's postmodification (*relative to aether*), the distance would be only 4 steps. Similarly, in Example 1, the distance between the direct object *velocity* and its verbal head *calculate* is 4 steps, which would be only two steps without the premodifiers *total* and *plasma*. Following the principle of Dependency Length Minimization, we therefore predict that for subjects, premodification is preferred, while objects and obliques display a preference for postmodification.

The **number of dependents** affects the distance to the head as well, again depending on the syntactic role: We expect subjects with a large number of modifiers to show a preference for premodification, while objects and obliques are expected to display a preference for postmodification. We also predict that a greater **length of the head noun** decreases the likelihood of pre-modification: Larger structures increase memory load, in which case post-modification as the less complex modification

type may be preferred to reduce overall complexity. Moreover, discourse status needs to be considered, for which we use **definiteness** as a proxy. We consider discourse-new information to be more complex than given information: In order to limit excess complexity, we therefore predict new NPs (here: NPs without a definite determiner) to show a smaller likelihood of premodification, the more compressed and complex alternative, than given NPs (NPs with a definite determiner). Examples 1 and 2 would fulfill this expectation: The discourse-given NP *the total plasma velocity* in Example 1 is premodified, while the discourse-new NP *velocity relative to aether* in Example 2 is postmodified.

Finally, we expect to see an effect of **time**: Due to the diachronic development of scientific English towards an efficient register with more and more compressed structures (Degaetano-Ortlieb and Teich, 2022), the likelihood of premodification should increase with the progression of time.

3 Dataset

We use the Royal Society Corpus (RSC; Fischer et al., 2020; Menzel et al., 2021), a diachronic corpus of English scientific writing. The full version 6 contains the *Philosophical Transactions and Proceedings of the Royal Society of London* from 1665 to 1995, with over 290 million tokens in more than 47,000 documents. The corpus was built in accordance to FAIR principles (Wilkinson et al., 2016), preprocessed using standard tools (Baron and Rayson, 2008; Schmid, 1995) and annotated with meta-data (Menzel et al., 2021). These include author, year of publication, text type (e.g. article, lecture, report, obituary), primary topic and journal (e.g. Series A - Mathematics and Physics, Series B - Biology).

We use version 6.0.3 which was parsed with the Python package *stanza* (Qi et al., 2020) and contains Universal Dependencies annotations (de Marneffe et al., 2021; Nivre et al., 2017). This version ("good sentences version") contains fewer tokens than the full version 6, since ungrammatical sentences or sentences which might be problematic for parsing (e.g. appendices, image titles, foreign language sentences) were not considered. Table 1 shows the composition of this corpus version over time.

Years	# Texts	# Tokens
1665-1699	1,312	2,194,828
1700-1749	1,674	2,895,445
1750-1799	1,806	5,037,372
1800-1849	2,709	7,001,970
1850-1899	5,502	12,923,443
1900-1949	6,879	21,014,576
1950-1996	20,413	78,142,577

Table 1: Composition of the Royal Society Corpus (version 6.0.3) over time.

4 Statistical Analysis

4.1 Preprocessing

For this study, we sampled 3,805 documents from the corpus. We used stratified sampling by publication year, meaning that the proportion of documents per year of the whole corpus was maintained in the sample.

Using a script written in Python (Van Rossum and Drake, 2009), version 3.10.15, we extracted the noun phrases from these documents. We consider only NPs headed by a common noun and only top-level NPs, i.e. NPs which are not embedded in other NPs. For these, we extracted the following linguistic features: head noun, number of dependencies of the head noun, syntactic role of the head noun, number of modifiers and modification type (premodification, postmodification, both). We also extracted the following metadata features of the noun or its context: text ID, sentence ID, head noun ID, publication year, author(s), text type and journal. This procedure resulted in information about 1,986,592 NPs.

For the statistical analysis, we filtered the data: First, we considered only NPs which possessed at least one modifier and were either pre- or post-modified, but not both. Determiners were not counted as modifiers, but as dependents of the head noun. We removed outliers which were most likely the result of parsing errors: NPs with more than 20 dependents, NPs with a distance to the head greater than 25, NPs with a head noun consisting of more than 20 characters. We also only focused on some syntactic roles since many roles were not attested frequently enough in our data (e.g. indirect objects, roots of a sentence). We considered nominal subjects, direct objects and oblique arguments. For the nominal subjects and the obliques, the various sub-categories (e.g. *oblique agent*) were subsumed into the overall category (e.g. *oblique*). We only

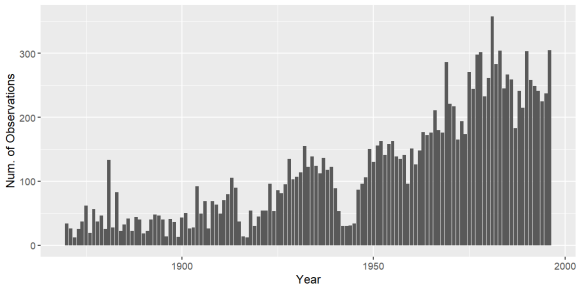

Figure 1: Number of observations per year.

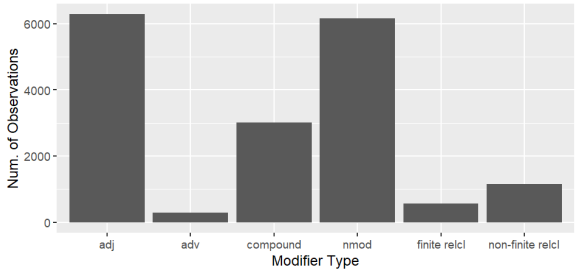

Figure 2: Number of observations of each modifier type.

considered the text type *article* in order to focus on scientific writing proper and to exclude non-scientific text types, such as obituaries and biographies. In addition, scientific text types other than *article* (e.g. lectures, speeches) were not strongly represented in the sample and contributed much fewer observations.

This filtering still resulted in 746,817 NPs, so we again applied stratified sampling by publication year, resulting in N = 14,934 observations to be included in the statistical analysis. The final dataset included 7,353 postmodified NPs and 7,581 premodified NPs. 7,843 NPs had no definite determiner, while 7,091 possessed a definite determiner. Most NPs (6,055) were oblique arguments, while 5,924 NPs acted as nominal subjects and 2,955 NPs as direct objects in their sentences. Most observations (5,605) stemmed from the journal *Proceedings of the Royal Society, Series A*, which encompasses the disciplines of mathematics, physics and engineering. An overview of the temporal distribution of our observations is given in Figure 1, with publication years ranging from from 1870 to 1996. The different types of modification in our data sample are shown in Figure 2.

4.2 Regression Model

We fit a mixed-effects logistic regression model in the statistical programming language R, version 4.4.2 (R Core Team, 2024) and using the library *glmmTMB* (Brooks et al., 2017). We chose regression modeling due to the large number of theoretically-motivated predictor variables which we took into account here: Mixed-effects regression modeling allows us to consider all of the variables and is appropriate given the hierarchical nature of corpus data and the resulting dependencies among observations (several observations from the same journal, author etc.).

Our dependent variable was modification type, with the levels *premodification* and *postmodification*. As predictor variables, we included year, distance to syntactic head, length of the head noun (in characters), discourse status (operationalized as the presence of a definite determiner for the status *Given*), sentence length (in number of words) and an interaction of dependency number and syntactic role. We also tested an interaction of dependency length and syntactic role, however, this model did not converge. The variable year was centered for ease of interpretation with regards to the intercept, all other numerical variables were centered and scaled. The factor variables were treatment-coded, with *postmodification* as the baseline for modification type and *nominal subject* as the baseline for syntactic role. To account for within-group variability, we included random intercepts for journal, author and noun, as well as a by-noun random slope for number of dependencies. Testing the variables for multicollinearity using the library *performance* (Lüdecke et al., 2021) revealed only mild correlation between the variables (variance inflation factors < 5). Model diagnostics (e.g. inspection of residuals) were performed with the package *DHARMa* (Hartig, 2024) and showed no overly problematic trends.

4.3 Results

The full model summary (Table 2) is included in Appendix A.

We found significant effects of year (p <0.001, z = 7.99, Figure 3), number of dependencies (p <0.001, z = 8.69, Figure 4), distance to syntactic head (p <0.001, z = -28.48, Figure 5), noun length (p <0.001, z = -6.67, Figure 6) and sentence length (p <0.001, z = -6.51, Figure 7): Over time, the likelihood of premodification over postmodification increases. The likelihood of premodification also increases for NPs with more dependencies. However, it decreases with greater distance to the head and with longer nouns and sentences.

We also found a significant effect of definiteness

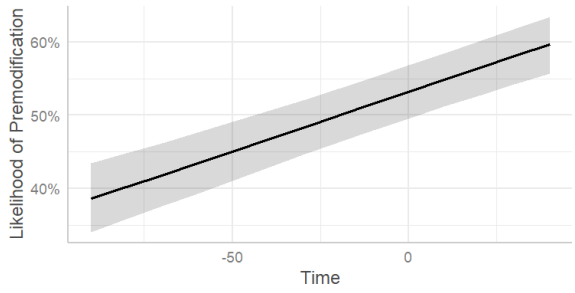

Figure 3: Effect of time on premodification likelihood: NPs in later years are more likely to be premodified.

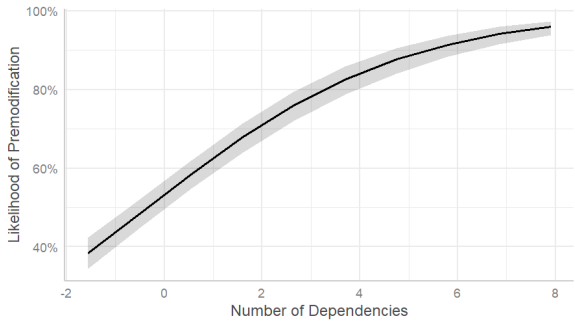

Figure 4: Effect of dependency number on premodification likelihood: NPs with more dependents are more likely to be premodified.

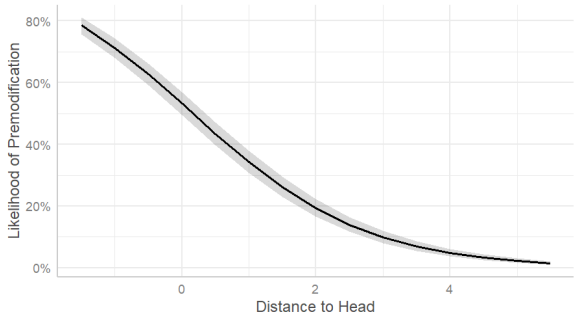

Figure 5: Effect of distance to head on premodification likelihood: NPs with greater distance to their head are less likely to be premodified.

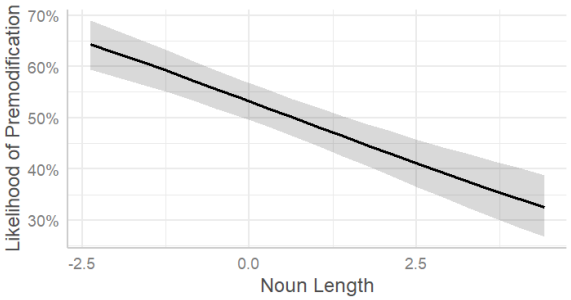

Figure 6: Effect of head noun length on premodification likelihood: NPs with a longer head noun are less likely to be premodified.

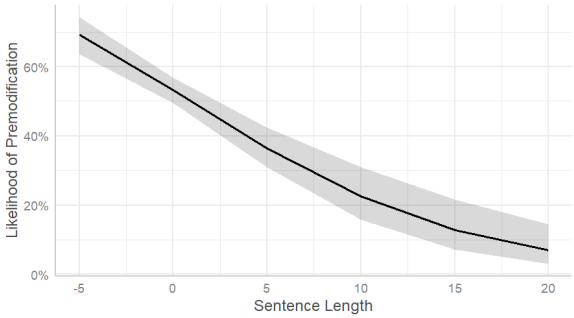

Figure 7: Effect of sentence length on premodification likelihood: NPs in longer sentences are less likely to be premodified.

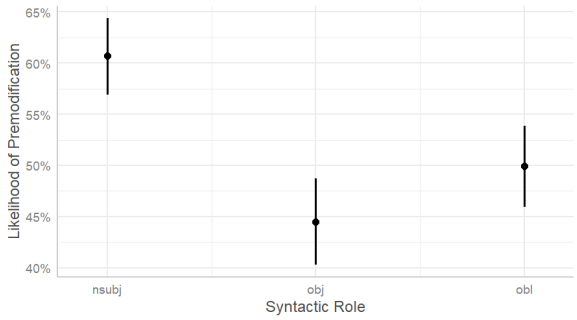

Figure 8: Effect of syntactic role on premodification likelihood: Subjects are the most likely to be premodified, followed by obliques and then direct objects.

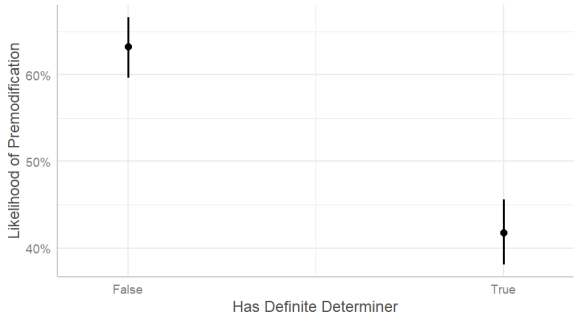

Figure 9: Effect of definiteness on premodification likelihood: Indefinite NPs are more likely to be premodified.

(p <0.001, z = -19.20, Figure 9) and of syntactic role (Figure 8): NPs without a definite determiner (i.e. NPs with either an indefinite or no determiner) had a higher chance of being premodified than noun phrases with a definite determiner. Compared to nominal subjects, NPs acting as direct objects (p <0.001, z = -9.94) or obliques (p <0.001, z = -7.65) had a lower chance of being premodified, with direct objects being the most unlikely to have premodification.

The interaction of dependency number and syntactic role, however, was not significant.

4.4 Discussion and Conclusion

Some of our predictions could be confirmed by the statistical analysis: In line with the principle of DLM, we observed a strong preference for subjects to be premodified, while direct objects and obliques were more likely to be postmodified (such as the object NP in Example 4). Moreover, a greater length of the head noun decreased the likelihood of premodification (consider the NP in Example 5 with a 7-syllable head noun). This supports the hypothesis that, in the face of higher memory load, language users opt for a less compressed modification type in order to ease processing difficulty. High memory load may also be an explanation for the observation that NPs in longer sentences are less likely to be premodified.

(4) Studies of these pterosaurs have revealed **a number of general features with regard to patterns of bone ossification** [...]. *(RSC, rspb_1996_0008)*

(5) **Reproducibility of results**, greater methodological awareness, and more rigorous assessment of hypothesis robustness are identified as additional issues [...]. *(RSC, rspb_1996_0205)*

Contrary to our expectations, however, definite NPs were less likely to be premodified than indefinite NPs. Our expectation may not have been confirmed because the notion of givenness was insufficiently operationalized by definiteness.

A larger distance to the verbal head was generally associated with a decreased likelihood of premodification. DLM can explain this for objects and obliques: In their case, premodification further increases the distance to the verbal head and should therefore be avoided. For subject NPs, a larger distance to their head in combination with postmodification (see Example 6) might be the result of an attempt to avoid the increased compression of premodified structures when memory load is high: If many dependents need to be integrated in the NP and stored in memory, writers might aim to avoid additional processing strain by selecting a less complex modification type.

(6) **Adoption of cladistic methods by students of archosaurs** has clearly been a slow and gradual process. *(RSC, rspb_1996_0205)*

It is interesting that the influence of dependency number was not modulated by syntactic role (non-significant interaction): Contrary to our expectations, NPs with more dependents were generally more likely to be premodified and not only in the case of subjects. This may be due to the influence of predictability: Some constituents of complex NPs might actually be very commonly used together and have a high transitional probability between the constituents. Encountering the first

element(s) of such an NP might lead the reader to correctly predict the whole structure. Over time, this facilitating effect for comprehension may lead to a preference in production. This may in particular be the case for compounds, highly compressed structures, which are derived from a process between syntax and morphology. Compound processing has been shown to be influenced by various factors such as constituent frequency, compound frequency, compound word length, compound family size or semantic transparency (Baayen et al., 2010; Schmidtke et al., 2021). Some factors actually have a facilitating effect on processing, so that compounds with high-frequency constituents and high semantic transparency are processed faster than compounds with low-frequency constituents and low semantic transparency. These effects might counteract and outweigh factors decreasing processing speed. A technical term consisting of an NP with several nominal modifiers, such as *heat shock cognate protein*, might be considered complex when judging merely from its syntactic structure: However, since *heat shock* and *protein* as well as *cognate* and *protein* co-occur frequently in biochemical texts and are established terms, this syntactic complexity might be outweighed by lexical frequency effects.

This study gives further support to the principle of Dependency Length Minimization and shows that it is also relevant for the choice between pre- and postmodification. It also supports the hypothesis that premodification might indeed be adding complexity to an NP, since it is dispreferred in the case of longer dependencies or for NPs with longer head nouns. However, this analysis also highlights that other factors, such as predictability and co-occurrence patterns, need to be considered as well when investigating optimization mechanisms in language use. Overall, this analysis supports the theory that optimization plays an important role in the evolution of scientific writing (Degaetano-Ortlieb and Teich, 2022). Considering the key role of NPs in scientific language, the results highlight the fact that these optimization pressures also act on the NP-level.

From a diachronic perspective, our study shows that the likelihood of premodification increases over time, even when controlling for other variables influencing the choice. This points towards a conventionalization trend within scientific writing: As register-specific norms become established over time, more compressed NPs are preferred, possibly outweighing competing constraints.

5 Limitations and Outlook

A major limitation of this analysis is the way discourse-given and discourse-new NPs were identified: While givenness and definiteness are correlated in English, they are not identical (Gundel et al., 1988): An NP with a demonstrative is usually given, an NP with a definite article, on the other hand, does not necessarily have to be given, only uniquely identifiable (Gundel et al., 1993). An investigation with more refined discourse annotation might lead to clearer insights on the influence of this factor.

Moreover, since our focus was on NPs headed by a common noun, other possible heads of NPs, like pronouns and proper names, were not included in this analysis. Future investigations should consider them as well in order to investigate if the results presented here are generalizable to pronouns and proper names. Special consideration should also be given to compounds, since the relationship between head and modifier(s) of a compound are presumably stronger than between head and phrasal modifiers.

NPs with both pre- and postmodification were also not considered here. It might be interesting to look at them in future research: Which dependents are added before and which after the head noun? Length and internal structure of modifiers are also a factor of interest, since modifiers may themselves contain heads with dependents. Investigating these aspects more closely may shed more light on the internal order of NP constituents and on the question whether the same principles apply here as for the clause level. It may also illustrate in more detail how competing pressures interact with each other in the process of language optimization.

Acknowledgements

The author would like to thank Elke Teich and Stefania Degaetano-Ortlieb as well as three anonymous reviewers for their insightful remarks on a previous version of this paper. This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102 Information Density and Linguistic Encoding.

References

Jennifer E. Arnold, Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. Heaviness vs. Newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76(1):28–55.

Harald Baayen, Victor Kuperman, and Raymond Bertram. 2010. Frequency effects in compound processing. In *Cross-disciplinary Issues in Compounding*, pages 257–270. John Benjamins Publishing Company.

Alan D. Baddeley, Neil Thomson, and Mary Buchanan. 1975. Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, 14(6):575–589.

David Banks. 2008. *The Development of Scientific Writing: Linguistic Features and Historical Context*. University of Toronto Press.

Alistair Baron and Paul Rayson. 2008. Vard2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*.

Otto Behaghel. 1909. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen*, 25:110.

Shaina Benjamin and Daniel Schmidtke. 2023. Conceptual combination during novel and existing compound word reading in context: A self-paced reading study. *Memory & Cognition*, 51(5):1170–1197.

Douglas Biber and Bethany Gray. 2011. Grammatical change in the noun phrase: The influence of written language use. *English Language & Linguistics*, 15(2):223–250.

Mollie E. Brooks, Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper W. Berg, Anders Nielsen, Hans J. Skaug, Martin Maechler, and Benjamin M. Bolker. 2017. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2):378–400.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Stefania Degaetano-Ortlieb. 2021. Measuring informativity: The rise of compounds as informationally dense structures in 20th-century scientific English. In *Corpus-based Approaches to Register Variation*, pages 291–312. John Benjamins Publishing Company.

Stefania Degaetano-Ortlieb and Elke Teich. 2022. Toward an optimal code for communication: The case of scientific English. *Corpus Linguistics and Linguistic Theory*, 18(1):175–207.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Matthias Eitelmann. 2016. Support for end-weight as a determinant of linguistic variation and change. *English Language & Linguistics*, 20(3):395–420.

Stefan Fischer, Jörg Knappen, Katrin Menzel, and Elke Teich. 2020. The Royal Society Corpus 6.0: Providing 300+ years of scientific writing for humanistic study. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 794–802. European Language Resources Association.

Linda S. Flower and John R. Hayes. 1981. A cognitive process theory of writing. *College Composition & Communication*, 32(4):365–387.

Richard Futrell. 2019. Information-theoretic locality properties of natural language. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 2–15.

Christina L. Gagné and Edward J. Shoben. 1997. Influence of thematic relations on the comprehension of modifier–noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(1):71.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson et al. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, Language, Brain*, 2000:95–126.

Dominic Guitard, Andrew J. Gabel, Jean Saint-Aubin, Aimée M. Surprenant, and Ian Neath. 2018. Word length, set size, and lexical factors: Re-examining what causes the word length effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(11):1824.

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 1988. On the generation and interpretation of demonstrative expressions. In *Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics*.

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, pages 274–307.

Michael A. K. Halliday. 1988. On the language of physical science. *Registers of written English: Situational factors and linguistic features*, 162:177.

Florian Hartig. 2024. *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. R package version 0.4.7.

John R. Hayes and Linda S. Flower. 1987. On the structure of the writing process. *Topics in Language Disorders*, 7(4):19–30.

Annie Jalbert, Ian Neath, Tamra J. Bireta, and Aimée M. Surprenant. 2011. When does length cause the word length effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(2):338.

Natalia Levshina. 2022. *Communicative Efficiency*. Cambridge University Press.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171–193.

Daniel Lüdecke, Mattan S. Ben-Shachar, Indrajeet Patil, Philip Waggoner, and Dominique Makowski. 2021. performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, 6(60):3139.

Katrin Menzel, Jörg Knappen, and Elke Teich. 2021. Generating linguistically relevant metadata for the Royal Society Corpus. *Research in Corpus Linguistics*, 9(1):1–18.

Sergio Mota and José Manuel Igoa. 2017. Parsing complex noun phrases: Effects of hierarchical structure and sentence position on memory load. *The Spanish Journal of Psychology*, 20:E37.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Diarmuid Ó Séaghdha. 2007. Designing and evaluating a semantic annotation scheme for compound nouns. In *Proc. Corpus Linguistics*, pages 1–17.

Ellen F. Prince. 1992. The ZPG letter: Subjects, definiteness, and information-status. In *Discourse Description: Diverse linguistic analyses of a fund-raising text*, pages 295–326. John Benjamins Publishing Company.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop*.

Daniel Schmidtke, Julie A. Van Dyke, and Victor Kuperman. 2021. Complex: An eye-movement database of compound word reading in english. *Behavior Research Methods*, 53:59–77.

David Temperley. 2007. Minimization of dependency length in written English. *Cognition*, 105(2):300–333.

Guido Van Rossum and Fred L. Drake. 2009. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):1–9.

A Appendix: Regression Model Summary

	Est.	**SE**	***z***	***p***
Intercept	0.85	8.48e-02	10.03	<0.001
Year	6.55e-03	8.20e-04	7.99	<0.001
Dependency Number	3.77e-01	4.34e-02	8.69	<0.001
Distance to (Verbal) Head	-7.80e-01	2.74e-02	-28.48	<0.001
Head Noun Length	-1.95e-01	2.92e-02	-6.72	<0.001
Syntactic Role *direct object*	-6.57e-01	6.61e-02	-9.94	<0.001
Syntactic Role *oblique*	-4.39e-01	5.74e-02	-7.65	<0.001
Definiteness	-8.73e-01	4.55e-02	-19.20	<0.001
Sentence Length	-1.37e-01	2.10e-02	-6.51	<0.001
Dep. Num * Synt. Role *direct obj.*	5.65e-02	6.79e-02	6.86e-01	0.493
Dep. Num * Synt. Role *oblique*	-7.69e-04	5.54e-02	-1.40e-02	0.989

Table 2: Regression model summary.