Exploring the Effect of Nominal Compound Structure in Scientific Texts on Reading Times of Experts and Novices

Anonymous ACL submission

Abstract

We explore how different types of nominal compound complexity in scientific writing, in particular different types of compound structure, affect the reading times of experts and novices. We consider both in-domain and out-of-domain reading and use PoTeC (Jakobi et al., 2024), a corpus containing eye-tracking data of German native speakers reading passages from scientific textbooks. Our results suggest that some compound types are associated with longer reading times and that experts may not only have an advantage while reading in-domain texts, but also while reading out-of-domain.

1 Introduction

Complex noun phrases (NPs), in particular nominal compounds, are used frequently in scientific writing and constitute a distinctive feature of this register (Biber and Gray, 2011). Nominal compounds allow for information to be transmitted in a highly compressed way, which increases implicitness (Biber and Gray, 2010): Logical relations between the constituents of a compound are implicit. Selecting a relational meaning from a range of possible meanings is therefore a crucial task in compound processing (Benjamin and Schmidtke, 2023). Possible meaning relations are in competition with each other, and compounds with a larger number of possible relations between constituents have been shown to pose a greater challenge for processing (ibid.). From a diachronic perspective, nominal compounds are a typical result of lexicalization processes in a language's morphological evolution (Hilpert, 2019). In the development of scientific writing, this process is especially productive due to ongoing terminology formation, which goes hand in hand with the increasing specialization of scientific disciplines: concepts are introduced to the community by using syntactically transparent renderings such as prepositional phrases or relative clauses (e.g. *methods that are used for the extraction of proteins*), and once they become established in the community they are compressed into less explicit renderings such as nominal compounds (e.g. *protein extraction methods*). A compound's successful processing can thus be assumed to rely on sufficient background knowledge to infer implicit relations between the compound's components. However, to our knowledge, there is no behavioral evidence for this assumption. While it is difficult to trace the establishment and processing of a compound over time within a scientific community, in the present study, we want to test whether background knowledge facilitates the processing of compounds differing in their internal complexity and structure. We model background knowledge as the reader's expertise in a scientific discipline. More specifically, we test whether in-domain experts and novices process compounds differently from out-of-domain experts and novices.

2 Background

Previous literature indicates that complexity on various linguistic levels can pose challenges in sentence processing. Syntactically more complex structures include longer dependencies between a syntactic head and its dependent, increasing their syntactic integration cost (cf. Dependency Locality Theory; Gibson, 1998). Specifically for nouns, (Demberg and Keller, 2008) have found that dependency locality predicts reading times for nouns. Other studies have considered word frequency and novelty as complexity features and found a correlation with increased reading times (e.g. Just and Carpenter, 1980, for scientific texts). Frequency-effects are also well known for the reading of compounds, with previous studies showing that higher constituent frequency, among other factors, ease processing (Baayen et al., 2010; Schmidtke et al., 2021). Likewise, the use of domain-specific ter-

minology (Škrjanec et al., 2023) has been found to influence reading time. In fact, having a distinctive code is beneficial for communication as transmission of information becomes more error-free (Harris, 1991).

Individual reader characteristics, such as background knowledge and experience have also been observed to influence reading comprehension (Kendeou and Van Den Broek, 2007). This is particularly relevant for scientific texts, which are targeted at an expert audience (Halliday, 1988). Over time, scientific language has shown to become more informationally dense with a tendency towards structural compression (Biber and Gray, 2013) and the use of dense phrasal structures (Halliday and Martin, 1993; Mair, 2006; Degaetano-Ortlieb and Teich, 2019). Mechanisms of specialization and conventionalization seem to act as balancing forces to modulate the transmission of information (Degaetano-Ortlieb and Teich, 2019). Specialization requires new forms of expression, given the need to express new concepts during periods of scientific innovation. Conventionalization allows for the formation of terminology known among experts, with compounds being the most compact forms of expression.

While previous studies considering compounds have often focused on English and mostly considered the prototypical compound structure noun-noun (e.g. Baayen et al., 2010; Schmidtke et al., 2021), our focus is on German and diverse types of compound structures (e.g., affix-adjective-noun-noun as in *Hyperfeinstrukturenaufspaltungen*, noun-affix-noun, such as *Cellulose-Mikrofibrillen*), assuming that different types of complexity impact their processing.

3 Hypotheses

Our hypotheses regarding the processing of different types of compound complexity are divided into two factors: length and structure. Regarding length, we assume that the more constituents a compound possesses, the more possible relations need to be inferred, making it harder to process. Regarding structure, we are interested in whether the parts-of-speech constituting the compound affect the compound's processing, i.e. noun-noun compounds vs. adjective-noun compounds. Noun-noun compounds might be easier to process due to their higher frequency. However, the meaning relation between the constituents of an adjective-noun compound can usually be described as "[head-noun] is [modifier-adjective]" (e.g., *blackbird*). Noun-noun compounds, on the other hand, possess more diverse meaning relations, such as "[head-noun] made from [modifier-noun]" (e.g., *olive oil*) or "[head-noun] for [modifier-noun]" (e.g., *baby oil*). This could make them harder to process than adjective-noun compounds.

Our two main hypotheses are as follows: (H1) Compounds differ in reading times given their internal structure, and (H2) expert knowledge influences reading times.

For H1, we will test the following hypotheses:

H1.1 Structurally more complex compounds, i.e. compounds with more constituents are harder to process and correlated with higher reading times.

H1.2 Compounds with non-nominal modifiers are processed differently than compounds with nominal modifiers, leading to a difference in reading times.

We also consider differences in compound processing based on reader characteristics (H2): We expect novices and out-of-domain readers to have more difficulty with compounds, since background knowledge plays an important role in inferring implicit relations. Additionally, experts are likely to outperform novices when reading texts from other scientific fields, as their general scientific reading competence provides an extra advantage. Our hypotheses regarding reader characteristics are therefore as follows:

H2.1 Compared to domain experts, novices and out-of-domain readers have generally more difficulties in compound processing and therefore longer reading times.

H2.2 When reading out-of-domain, experts still have fewer difficulties in compound processing than novices, and therefore shorter reading times.

The results can highlight the impact of NP complexity on processing difficulty and its interaction with readers' domain expertise. Besides being of theoretical interest, these findings are relevant for teaching English for Academic Purposes. Studies like Priven (2020) suggest that non-native English speaking students experience difficulties in understanding complex noun phrases in academic

writing. Gaining a better insight into which structures are particularly challenging may guide future teaching.

4 Data and Preprocessing

We use PoTeC (Jakobi et al., 2024), a German naturalistic eye-tracking-while-reading corpus. It contains the data of 75 German native speakers who were university students of either biology or physics. The students were either experts (graduate students) or novices (undergraduate students) and read passages from biology and physics textbooks. The corpus contains various reading time measures (e.g., first-pass reading time, total fixation time, number of incoming regressions, number of outgoing regressions) and linguistic annotation (e.g., part of speech, frequency, surprisal estimates from different language models).

The corpus also contains dependency annotation and constituency annotation based on the Python library *spacy* (Honnibal and Montani, 2017). In order to get a more fine-grained dependency annotation based on Universal Dependencies (Nivre et al., 2017), we parsed and annotated the corpus files with the help of the Python library *stanza* (Qi et al., 2020). Since compounds written as one word (which is the case for most German compounds) are not specifically annotated under this scheme and compounds separated by a hyphen are only superficially annotated, we then extracted all the nouns, manually identified the compounds and annotated them: For each compound, we identified its constituents and annotated their part of speech. In the case of neo-classical compounds, i.e. compounds containing a constituent originating from Latin or Greek, the part of speech could not be clearly identified. We used the tag *affix* here, in accordance with German dictionary conventions. The compounds were labeled by one annotator, annotations were subsequently validated by another person. In the case of disagreements, a third person was consulted. Table 1 shows some examples of our annotation.

Table 2 shows the total number of observations and the number of unique compound words per compound category, for biology and physics respectively. For both domains, most compounds belonged to the *noun-noun* category, which is the prototypical compound in German (see also studies regarding first language acquisition, e.g., Korecky-Kröll et al., 2017).

In addition, information about the number of occurrences was added for each compound, since many compounds occurred several times in the stimulus texts: The first occurrence of a specific compound was labeled as *1*, subsequent occurrences as *2*, *3* and so on. We also included information about the first constituent frequency, since constituent frequency effects for compounds are well known in the literature. The first constituent frequencies were extracted from the *dlexDB* database (Kliegl et al., 2025), a reference database for German which was also used in the creation of PoTeC.

5 Influence of Constituent Number

For our first analysis, we consider the influence of constituent number (H1.1). More specifically, we investigate whether compounds with two constituents are read faster than compounds with three constituents. We also investigate the influence of background knowledge (H2.1 and H2.2). For this, we conducted an analysis on biology texts and another on physics texts to study in-domain vs. out-of-domain reading behavior. For biology, we analyzed N = 4984 observations (first-pass reading times of individual compounds): Of these observations, 4261 were compounds with two constituents, 723 compounds had three constituents. For physics, we analyzed N = 4681 observations, including 4256 observations with two constituents and 425 observations with three constituents. We only considered compounds that were fixated at least once and which were fixated during first-pass reading. We also excluded compound words that occurred in sentence-initial position and for which no first constituent frequency could be retrieved from the reference database.

5.1 Regression Model

For each domain, we fit generalized mixed effects regression models using the *glmmTMB* package (Brooks et al., 2017) in the statistical programming language R, version 4.4.2 (R Core Team, 2024). Our dependent variable was first-pass reading time. Since reading times, like other reaction time data, are not normally distributed (Lo and Andrews, 2015), we used gamma regression models with a log-link. Using gamma models for reaction time data has been suggested in the literature as a possible alternative to log-transforming the data before analysis, which is considered to be problematic by some authors (Lo and Andrews, 2015).

Compound	English Translation	Division	Word Class
Hyperfeinstrukturenaufspaltungen	hyperfine structure splitting	Hyper-fein-strukturen-aufspaltungen	affix-adjective-noun-noun
Gelelektrophorese	gel electrophoresis	Gel-elektro-phorese	noun-affix-noun
Cellulose-Mikrofibrillen	cellulose microfibrils	Cellulose-Mikro-fibrillen	noun-affix-noun
Prionenprotein	prion protein	Prionen-protein	noun-noun

Table 1: Compound annotation with English equivalents, division, and word class structure.

Category	Biology		Physics	
	Total	Unique	Total	Unique
adj-n	375	5	525	6
adj-n-n	0	0	150	2
adj-n-n-n	75	1	0	0
aff-adj-n-n	0	0	75	1
aff-aff-n	75	1	75	1
aff-n	450	5	525	5
aff-n-n	75	1	150	2
adv-n	300	2	0	0
n-aff-n	150	2	0	0
n-n	3900	41	3375	36
n-n-n	450	5	75	1
n-n-n-n	225	1	0	0
v-n	0	0	150	2

Table 2: Compound category counts in Biology and Physics, with total and unique counts.

Our predictors of interest were the interaction of compound structure and domain expert status and the interaction of technicality and domain expert status. The factor compound structure had the levels *two constituents* and *three constituents*. The factor technicality had the levels *technical* and *non-technical*. The levels of domain expert status were *novice biologist*, *expert biologist*, *novice physicist*, *expert physicist*. For the biology texts, the biologists were reading in-domain and the physicists were reading out-of-domain. For physics texts, it was vice versa. In this way, we model the compound structure while taking into account the reader's level of expertise and domain familiarity.

We controlled for word length, type frequency of the whole compound, lemma frequency of the first constituent, surprisal (i.e., word predictability in context; Shannon, 1948), word index in the sentence, hyphenation and occurrence number of the compound word, since many compounds occurred more than once in the stimulus texts. Our control variables were theoretically motivated, based on factors known to influence reading behavior (see Section 2). We opted not to use step-wise model selection due to concerns about the generalizability of the resulting model (see, e.g., Smith, 2018). Finally, we included by-subject and by-lemma random intercepts and a by-lemma random slope for surprisal to account for subject- and lemma-based variability. The factors compound structure, domain expert status, technicality and hyphenation were treatment-coded, with two constituent compounds, domain expert, non-technical term and non-hyphenated compound as the baseline levels. The frequency-based variables were log-transformed, while the variable word length was centered and scaled.

For model diagnostics, we inspected the residuals using the R package *DHARMa* (Hartig, 2024). The plots did not show any overly problematic trends. We also tested for collinearity using the package *performance* (Lüdecke et al., 2021): Overall collinearity was low, with variance inflation factors below 2.

5.2 Results

The significant results ($\alpha = 0.05$) for biology are shown in Table 3. The full model summary is included in the appendix (note that the model coefficients are on the log-scale).

	Est.	SE	z	p
Intercept	6.06	0.10	58.31	<0.001
word length	0.18	0.03	5.32	<0.001
surprisal	0.02	0.00	4.20	<0.001
word index	-0.01	0.00	-3.27	<0.01
novice physicist, technical term	0.26	0.05	5.06	<0.001
expert physicist, technical term	0.22	0.04	5.02	<0.001

Table 3: Analysis of constituent number: significant coefficients for biology.

We observed a significant interaction of technicality and domain expert status for novice physicists ($\beta = 0.26$, SE = 0.05, $p < 0.001$) and expert physicists ($\beta = 0.22$, SE = 0.04, $p < 0.001$), i.e. out-of-domain readers when reading technical terms.

	Est.	SE	z	p
Intercept	6.06	0.15	41.54	<0.001
word length	0.10	0.04	2.65	0.008
compound frequency	-0.15	0.06	-2.33	0.02
word index	0.00	0.00	2.07	0.04
hyphenation	-0.46	0.20	-2.37	0.02
novice biologist, technical term	0.10	0.05	2.22	0.03
expert biologist, technical term	0.09	0.04	2.25	0.02
novice biologist, three constituents	0.19	0.08	2.44	0.01

Table 4: Analysis of constituent number: significant coefficients for physics.

Figure 1 shows the predicted reading times for non-technical vs. technical terms and for the different reader groups: While the reading times for technical terms are generally higher than for non-technical terms, and while out-of-domain readers are generally slower than in-domain readers, out-of-domain readers are particularly slow when reading technical terms. This holds for both novice and expert physicists, with novice physicists showing a slightly larger increase in reading times.

The effects of our control variables have been attested in previous studies. We observed significant effects of word length (β = 0.18, SE = 0.03, p < 0.001), surprisal (β = 0.02, SE = 0.00, p < 0.001) and word index in sentence (β = -0.01, SE = 0.00, p < 0.01): Longer words and words with higher surprisal were associated with increased reading times, while words with a higher index (i.e. a later position) in the sentence were associated with decreased reading times.

The significant effects (α = 0.05) for physics are shown in Table 4 (see complete model summary in the appendix).

We found a significant effect of compound structure when the reader was a novice biologist and the compound consisted of three constituents (β = 0.19, SE = 0.08, p < 0.05). The reading times associated with compounds with three constituents were generally higher than for those with two constituents. This effect was statistically significant for novice biologists, who showed longer reading times compared to expert physicists reading two-constituent compounds. Model predictions for this interaction are shown in Figure 2.

In addition, there was a significant interaction of domain expert status and terminology for novice biologists (β = 0.10, SE = 0.05, p < 0.05) and expert biologists (β = 0.09, SE = 0.04, p < 0.05): Both groups show increased reading times for technical terms, compared to expert physicists reading non-technical terms. The increase is slightly higher for the novice biologists.

We also observed an effect of the control variables word length (β = 0.10, SE = 0.04, p < 0.01), compound frequency (β = -0.15, SE = 0.06, p < 0.05), word index (β = 0.00, SE = 0.00, p < 0.05) and hyphenation (β = -0.46, SE = 0.20, p < 0.05). For word length and word index, the effect was similar to the one observed for the biology texts. Additionally, more frequent compounds and compounds containing a hyphen were read faster.

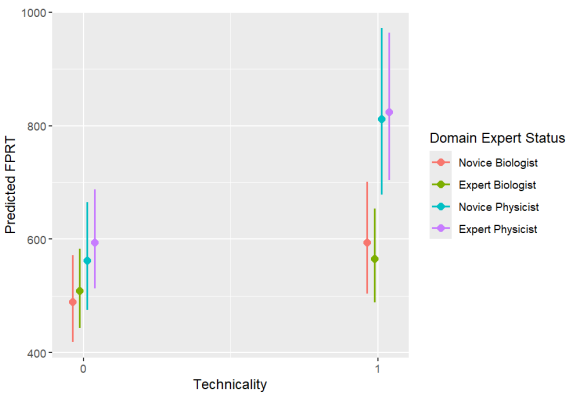

Figure 1: Biology: Predicted reading times for non-technical vs. technical terms.

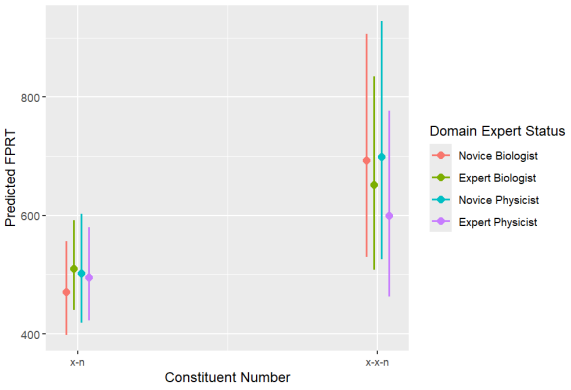

Figure 2: Physics: Predicted reading times for two- vs. three-constituent compounds

5.3 Discussion

Our results suggest an effect of compound structure on compound processing (H1.1), at least for the physics texts: Compounds with three constituents were generally read slower than compounds with two constituents, even when controlling for word length as we did in our model (note that compounds

with three constituents do not necessarily need to be longer than compounds with two constituents). This interacted with reader domain knowledge: Readers reading out-of-domain and possessing little expertise in their own field (novice biologists) showed a significant increase in reading times for three-constituent compounds. Expert biologists, on the other hand, seemed to have fewer difficulties, since they did not diverge that significantly from in-domain experts. This might again indicate a general scientific reading skill providing them with an advantage.

In addition, we found evidence that technicality may have an effect on reading times and that this varies by reader expertise: For biology texts, out-of-domain readers were particularly slow when reading technical compounds, reflecting processing difficulties due to their lack of familiarity with the subject matter. The slightly greater increase in reading times for novice physicists compared to expert physicists also suggests that experts may indeed still have an advantage when reading texts from a different domain. The results for biology, therefore, support H 2.1 and H 2.2. For physics texts, the picture was similar: Out-of-domain readers generally showed increased reading times for technical compounds. The increase was slightly higher for novice biologists than for expert biologists, suggesting an expert advantage even when reading out-of-domain.

Moreover, our analysis showed the expected effects of some well-known factors influencing compound processing: greater word length and higher surprisal were associated with increased reading times. A later position of the compound in the sentence, higher frequency and hyphenation, on the other hand, were associated with decreased reading times.

6 Influence of Modifier Type

For our second analysis, we now considered the influence of modifier type (H1.2). Extracting all two-constituent compounds, we compared the prototypical *noun-noun* compounds with those compounds in which the modifier has a different word class, e.g., *verb-noun* or *adjective-noun*. In total, this led to $N = 4261$ observations to be analyzed for biology. 3408 observations were *noun-noun* compounds, 853 observations were compounds with a non-nominal modifier. For physics, we analyzed 4256 observations: 3147 *noun-noun* compounds and 1109 compounds with a non-nominal modifier.

	Est.	SE	z	p
Intercept	6.00	0.11	53.67	<0.001
word length	0.18	0.05	3.92	<0.001
surprisal	0.02	0.01	4.07	<0.001
word index	-0.01	0.00	-3.19	0.001
expert status: expert physicist	0.17	0.07	2.47	0.01
novice physicist, technical term	0.25	0.05	4.69	<0.001
expert physicist, technical term	0.22	0.05	4.83	<0.001
novice physicist, non-nom. mod.	-0.16	0.06	-2.56	0.01

Table 5: Analysis of modifier type: significant coefficients for biology.

6.1 Regression Model

We fit generalized linear mixed-effects models in the same way as in Section 5, with the exception of the predictor compound type, which now consisted of the levels *noun-noun* and *other-noun*. Again, the factor compound type was treatment coded, with *noun-noun* as the baseline level.

Inspecting the model residuals revealed no overly problematic trends. The collinearity of our predictors was moderate to low, with variance inflation factors below 3 for the biology model and below 2 for the physics model.

6.2 Results

The significant effects ($\alpha = 0.05$) for biology are displayed in Table 5. The full model summary is included in the appendix.

We observed an effect of modifier type and reader background on reading times (β = -0.16, SE = 0.06, $p < 0.05$): Out-of-domain readers with little experience in their own field (novice physicists) diverge significantly from expert biologists. Interestingly, they have shorter reading times for compounds with non-nominal modifiers. We will return to this point in the discussion. Model predictions for compound type are displayed in Figure 3.

In addition, we see a significant interaction of technicality and reader expertise: Similarly to the results from Section 5, out-of-domain readers, namely novice ($\beta = 0.25$, SE = 0.05, $p < 0.001$) and expert physicists ($\beta = 0.22$, SE = 0.05, $p < 0.0001$) are relatively slow when reading technical compounds. The increase in reading times was slightly higher for the novice physicists.

	Est.	SE	z	p
Intercept	6.11	0.15	38.84	<0.001
word length	0.08	0.04	2.20	0.03
compound frequency	-0.14	0.07	-2.08	0.04
hyphenation	-0.63	0.26	-2.40	0.02
novice biologist, technical term	0.14	0.05	2.87	<0.01
expert biologist, technical term	0.12	0.04	2.73	<0.01

Table 6: Analysis of modifier type: significant coefficients for physics.

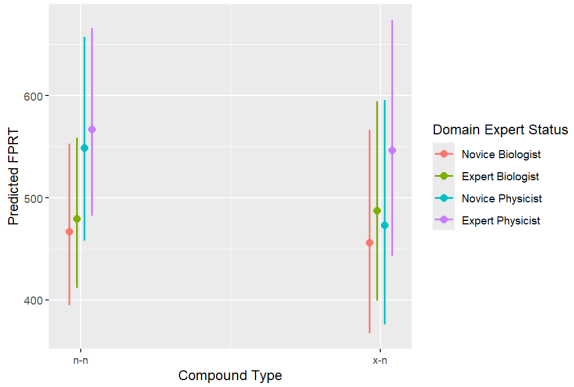

Figure 3: Biology: Predicted reading times for compounds with a nominal vs. non-nominal modifier.

We also observed significant effects of the control variables word length (β = 0.18, SE = 0.05, p < 0.001), surprisal (β = 0.02, SE = 0.01, p < 0.001), and word index in sentence (β = -0.01, SE = 0.00, p < 0.001): Longer and less predictable words were associated with increased reading times, while words occurring later in the sentence were read faster.

The significant effects (α = 0.05) for physics are displayed in Table 6. As before, the full model summary can be found in the appendix.

Similarly to the results in Section 5, out-of-domain readers, the novice (β = 0.14, SE = 0.05, p < 0.01) and expert biologists (β = 0.12, SE = 0.04, p < 0.01) diverge significantly from expert physicists in their reading behavior. Both groups have increased reading times, with a slightly higher increase for the novices.

The significant effects of our control variables existed for word length (β = 0.08, SE = 0.04, p < 0.05), compound frequency (β = -0.14, SE = 0.07, p < 0.05) and hyphenation (β = -0.63, SE = 0.26, p < 0.05): Reading times were higher for longer words, while more frequent as well as hyphenated compounds were associated with decreased reading times.

6.3 Discussion

Regarding the effect of technicality and reader background, the results of our second analysis yielded similar results as the analysis in Section 5: Again, readers with no background in the domain at hand were significantly slower for technical terms. The increase was larger for the novices than for the experts reading out-of-domain texts. This comes as no surprise since the data was roughly the same as in the previous analysis, only the factor compound type was coded differently. In our second analysis, we observed an effect of modifier type in the biology domain: Novice physicists diverged significantly from expert biologists and had shorter reading times for compounds with non-nominal modifiers compared to compounds with nominal modifiers. This supports hypothesis H1.2, indicating an effect of modifier type for processing. Interestingly, non-nominal modifiers may be easier to process: This might reflect the smaller number of possible semantic relations between head and modifier for, e.g., *adjective-noun* compounds compared to *noun-noun* compounds.

7 Discussion and Conclusion

In our two analyses, we saw some evidence supporting our initial hypotheses: Compound structure seemed to have an effect on reading time, suggesting differences in processing difficulty for compounds with different numbers of constituents and for compounds with different types of modifiers. However, this effect varied based on reader background: Novice biologists showed an increase of reading times for compounds with more constituents when reading texts from the physics domain. Novice physicists showed a decrease of reading times for compounds with non-nominal modifiers when reading texts from the biology domain. The fact that the effect of compound structure could only be observed for novice readers reading out-of-domain texts suggests that the effect might be relatively small and interacting with reader background: In our dataset, we could only observe it for readers with neither domain knowledge nor much experience in their own field. It also suggests that experts possess general scientific reading competence which helps them even when reading out-of-domain: They performed more similarly to in-domain readers even when reading texts from a dif-

ferent domain. The effect was only visible in some text domains: The effect of constituent number was only visible for the physics texts, while the effect of modifier type was only visible for the biology texts. Further studies would need to investigate the reasons for this difference and consider other domains and readers with other backgrounds. As natural sciences, biology and physics still have many similarities in their respective domain-specific lexicon. Effects of compound structure in out-of-domain readers might be more pronounced for readers with background in a more distant field (e.g., readers with a social science background reading physics or biology texts).

The effect of technicality and reader domain was relatively robust: Out-of-domain readers always had significantly longer reading times for technical terms than in-domain readers. For the out-of-domain readers, the experts showed a smaller increase in reading times, supporting the hypothesis of their general scientific reading competence.

Our analysis has one major limitation: The dataset was unbalanced, since most unique compounds belonged to the *noun-noun* category. The categories of compounds with three constituents and compounds with non-nominal modifiers contained far less unique words. Thus, the question remains if our significant effects can be attributed to idiosyncrasies of these individual compounds or if they can be generalized. Moreover, some categories were quite diverse internally: Non-nominal modifiers, for instance, encompassed different word classes which may not have the same effect on compound processing. An *adjective-noun* compound might pose different challenges than a *verb-noun* compound. For this reason, the current study could be replicated with a different dataset: Data with less imbalance in the compound categories might provide clearer results regarding the effect of compound structure and might allow a more fine-grained analysis. There are also additional variables to be considered in future research: the number of possible relations between constituents, compound transparency or constituent family size.

This would shed more light on the mechanisms of compound processing, in particular for compounds with more than two constituents and non-nominal modifiers. It would also enable us to gain more insights into the effect of reader knowledge on the processing of complex syntactic structures.

References

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.

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. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, 9(1):2–20.

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.

Douglas Biber and Bethany Gray. 2013. Nominalizing the verb phrase in academic science writing. In Bas Aarts, Joanne Close, Geoffrey Leech, and Sean Wallis, editors, *The Verb Phrase in English: Investigating Recent Language Change with Corpora*, Studies in English Language, pages 99–132. Cambridge University Press, Cambridge.

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.

Stefania Degaetano-Ortlieb and Elke Teich. 2019. Toward an optimal code for communication: The case of scientific english. *Corpus Linguistics and Linguistic Theory*, 0:1–33.

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

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

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

Michael A. K. Halliday and James R. Martin. 1993. *Writing Science: Literacy and Discursive Power*. Falmer Press, London.

Zellig Harris. 1991. *A theory of language and information. A mathematical approach*. Clarendon Pess, Oxford.

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

Martin Hilpert. 2019. Lexicalization in morphology. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Deborah N Jakobi, Thomas Kern, David R Reich, Patrick Haller, and Lena A Jäger. 2024. Potec: A german naturalistic eye-tracking-while-reading corpus. *arXiv preprint arXiv:2403.00506*.

Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: from eye fixations to comprehension. *Psychological review*, 87(4):329.

Panayiota Kendeou and Paul Van Den Broek. 2007. The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts. *Memory & cognition*, 35(7):1567–1577.

Reinhold Kliegl, Thomas Hanneforth, Alexander Geyken, Kay-Michael Würzner, Julian Heister, Edmund Pohl, Johannes Bubenzer, and Frank Wiegand. 2025. dlexdb – annotated lexical data.

Katharina Korecky-Kröll, Sabine Sommer-Lolei, and Wolfgang U Dressler. 2017. Emergence and early development of german compounds. In *Nominal compound acquisition*, pages 19–37. John Benjamins Publishing Company.

Steson Lo and Sally Andrews. 2015. To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in psychology*, 6:1171.

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.

Christian Mair. 2006. *Twentieth-Century English: History, Variation and Standardization*. Cambridge University Press, Cambridge.

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.

Dmitri Priven. 2020. "all these nouns together just don't make sense!": An investigation of eap students' challenges with complex noun phrases in first-year college-level textbooks. *Canadian Journal of Applied Linguistics*, 23(1):93–116.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

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

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.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Iza Škrjanec, Frederik Yannick Broy, and Vera Demberg. 2023. Expert-adapted language models improve the fit to reading times. *Procedia Computer Science*, 225:3488–3497.

G Smith. 2018. Step away from stepwise. *Journal of Big Data*, 5:32.

A Appendix: Regression Model summaries

	Est.	SE	z	p
Intercept	6.06	0.10	58.31	<0.001
compound: three constituents	0.10	0.11	0.88	0.39
word length	0.18	0.03	5.32	<0.001
compound frequency	-0.09	0.07	-1.23	0.22
surprisal	0.02	0.00	4.20	<0.001
word index	-0.01	0.00	-3.27	<0.01
hyphenation	0.02	0.10	-0.17	0.87
occurrence	0.01	0.02	0.29	0.77
first constituent frequency	0.01	0.02	0.49	0.62
expert status: novice biologist	-0.04	0.07	-0.53	0.59
expert status: novice physicist	0.10	0.08	1.25	0.21
expert status: expert physicist	0.16	0.07	2.28	0.23
technical term	0.11	0.09	1.66	1.21
novice biologist, technical term	0.09	0.05	1.90	0.06
novice physicist, technical term	0.26	0.05	5.06	<0.001
expert physicist, technical term	0.22	0.04	5.02	<0.001
novice biologist, three constituents	0.02	0.07	0.26	0.79
novice physicist, three constituents	0.04	0.07	0.50	0.62
expert physicist, three constituents	-0.02	0.06	-0.28	0.78

Table 7: Analysis of constituent number: model summary for biology. (Note that coefficients are on the log-scale.)

	Est.	SE	z	p
Intercept	6.06	0.15	41.54	<0.001
compound: three constituents	0.19	0.12	1.60	0.10
word length	0.10	0.04	2.65	0.008
compound frequency	-0.15	0.06	-2.33	0.02
surprisal	0.01	0.01	1.82	0.07
word index	0.00	0.00	2.07	0.04
hyphenation	-0.46	0.20	-2.37	0.02
occurrence	-0.01	0.02	-0.41	0.68
first constituent frequency	-0.01	0.02	-0.26	0.79
expert status: novice biologist	-0.05	0.09	-0.59	0.56
expert status: expert biologist	0.03	0.07	0.41	0.68
expert status: novice physicist	0.01	0.09	0.16	0.88
technical term	-0.03	0.09	-0.40	0.69
novice biologist, technical term	0.10	0.05	2.22	0.03
expert biologist, technical term	0.09	0.04	2.25	0.02
novice physicist, technical term	0.08	0.05	1.51	0.13
novice biologist, three constituents	0.19	0.08	2.44	0.01
expert biologist, three constituents	0.05	0.07	0.74	0.46
novice physicist, three constituents	0.14	0.09	1.58	0.11

Table 8: Analysis of constituent number: model summary for physics. (Note that coefficients are on the log-scale.)

	Est.	SE	z	p
Intercept	6.00	0.11	53.67	<0.001
compound: non-nominal mod.	0.02	0.11	0.14	0.89
word length	0.18	0.05	3.92	<0.001
compound frequency	-0.07	0.09	-0.76	0.45
surprisal	0.02	0.01	4.07	<0.001
word index	-0.01	0.00	-3.19	0.001
hyphenation	0.02	0.11	0.21	0.83
occurrence	0.01	0.02	0.38	0.70
first constituent frequency	0.01	0.03	0.41	0.69
expert status: novice biologist	-0.03	0.07	-0.36	0.72
expert status: novice physicist	0.13	0.08	1.68	0.09
expert status: expert physicist	0.17	0.07	2.47	0.01
technical term	0.08	0.10	0.77	0.44
novice biologist, technical term	0.08	0.05	1.60	0.11
novice physicist, technical term	0.25	0.05	4.69	<0.001
expert physicist, technical term	0.22	0.05	4.83	<0.001
novice biologist, non-nominal mod.	-0.04	0.06	-0.69	0.49
novice physicist, non-nominal mod.	-0.16	0.06	-2.56	0.01
expert physicist, non-nominal mod.	-0.05	0.05	-0.96	0.33

Table 9: Analysis of modifier type: model summary for biology. (Note that coefficients are on the log-scale.)

	Est.	SE	z	p
Intercept	6.11	0.15	38.84	<0.001
compound: non-nominal mod.	-0.08	0.11	-0.78	0.44
word length	0.08	0.04	2.20	0.03
compound frequency	-0.14	0.07	-2.08	0.04
surprisal	0.01	0.01	1.60	0.11
word index	0.00	0.00	1.96	0.05
hyphenation	-0.63	0.26	-2.40	0.02
occurrence	-0.01	0.02	-0.32	0.75
first constituent frequency	-0.01	0.03	-0.55	0.58
expert status: novice biologist	-0.09	0.09	-1.00	0.32
expert status: expert biologist	-0.00	0.08	-0.02	0.99
expert status: novice physicist	0.03	0.10	0.31	0.76
technical term	-0.06	0.10	-0.58	0.56
novice biologist, technical term	0.14	0.05	2.87	<0.01
expert biologist, technical term	0.12	0.04	2.73	<0.01
novice physicist, technical term	0.06	0.05	1.11	0.27
novice biologist, non-nominal mod.	0.07	0.06	1.19	0.24
expert biologist, non-nominal mod.	0.07	0.05	1.36	0.17
novice physicist, non-nominal mod.	-0.03	0.06	-0.43	0.67

Table 10: Analysis of modifier type: model summary for physics. (Note that coefficients are on the log-scale.)