Nominal Compounds, Information, and Scientific Texts

Vom Fachbereich
Sozialwissenschaften der
Rheinland-Pfälzischen Technischen Universität Kaiserslautern-Landau

zur Verleihung des akademischen Grades
Doktor rerum naturalium (Dr.rer.nat.)
genehmigte
Dissertation

von

John Cristian Borges Gambôa

Tag der Disputation:	Kaiserslautern, 17. April 2026
Dekanin:	Prof. Dr. Shanley E.M. Allen
Vorsitzender:	Prof. Dr. Thomas Schmidt
Gutachter:	1. Prof. Dr. Shanley E.M. Allen
	2. Prof. Dr. Elke Teich
Disputanten:	Prof. Dr. Maria Klatte
	Prof. Dr. James R. Anglin

DE 386

Kaiserslautern, Dezember, 2025

Nominal Compounds, Information, and Scientific Texts

PhD Thesis

John Cristian Borges Gambôa

RPTU University of Kaiserslautern-Landau, Germany
Kaiserslautern, December, 2025

Selbständigkeitsversicherung

Hiermit versichere ich,

- dass ich die vorgelegte Arbeit selbst angefertigt und alle benutzten Hilfsmittel in der Arbeit angegeben habe,
- dass ich diese Dissertation nicht schon als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht, und
- dass weder die gleiche noch eine andere Abhandlung der Dissertation bei einer anderen Universität oder einem anderen Fachbereich der Technischen Universität Kaiserslautern veröffentlicht wurde.

15. Dezember 2025, Kaiserslautern

For my mother, Lourdes Schmitz Borges

For my uncle, Daniel Schmitz Borges

For my grandmother, Lourdes Schmitz Borges

Contents

Acknowledgements **xi**

Abstract **xiii**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Structure . . . 5

2 Compounding and Nominal Compounds **7**
2.1 Open and closed classes of words . . . 8
2.2 Creating new words in a language . . . 9
2.3 A note on inconsistent terminology . . . 10
2.4 Noun Compounds . . . 16
2.4.1 What are noun compounds? . . . 16
2.4.2 Structural ambiguity . . . 17
2.4.3 The head noun and its position . . . 18
2.4.4 Relational ambiguity: how are a compound's words linked together? . . . 19
2.4.5 The scope of this thesis . . . 21
2.5 Previous research on CNC processing . . . 21
2.6 Why are CNCs used in scientific papers? . . . 25
2.7 Summary . . . 26

3 Information **27**
3.1 A gentle introduction to Information Theory . . . 28
3.1.1 Communication: The typical Information Theory problem . . . 28
3.1.2 A measure of information . . . 31
3.1.3 Entropy . . . 34
3.1.4 Mutual information and the channel capacity . . . 35
3.2 Applying these ideas to Psycholinguistics . . . 37
3.2.1 A possible psycholinguistic instantiation . . . 37
3.2.2 What does this mean for CNCs? . . . 38
3.3 The Entropy Rate Constancy (ERC) Principle . . . 39
3.4 The Uniform Information Density (UID) Hypothesis . . . 41
3.5 Summary . . . 42

4 Predictions **43**
4.1 RQ1: predictions for CNC processing . . . 43
4.2 RQ2: predictions for CNC use . . . 45

5 Publications on CNC processing **47**
5.1 CNC processing in the L1 . . . 47
5.2 CNC processing in the L2 . . . 48

6 Publications on CNC use **153**
6.1 How CNCs are set up in scientific papers . . . 153
6.2 The role of context in the perceived difficulty of CNCs . . . 154

7 General Discussion **221**
7.1 RQ1: the evidence related to CNC processing 221
7.2 RQ2: the evidence related to CNC use 223

8 Conclusion **227**

Bibliography **229**

Acknowledgements

This PhD was long. A long quest, a journey, with "hills" and "valleys", that culminated in the writing of this dissertation. Through the years that went by, and the pandemic that occurred in between, a quite large number of people have contributed, some only briefly, some throughout the whole endeavor, not only with the PhD itself (with suggestions on my writing, with the experiments, discussing results or providing guidance on what to do in certain bureaucratic situations), but also with maintaining the mental stability I so desperately needed in order to progress.

I would not have managed to complete this PhD without guidance, and thankfully, I had no shortage of people to resort to whenever in need. For that, I would like to thank Professors Maria Klatte, Thomas Lachmann, Thomas Schmidt, and Juhani Järvikivi, all of whom have contributed in their own ways to my understanding of what it means to be[come] a researcher. In this regard, I would also like to thank Leigh Fernandez, who patiently shared the office with me all these years, fabulously making us all drink some sparkling wine every few months; and, of course, my advisor Shanley Allen, to whom I feel profoundly obliged for providing me with the opportunity of moving into the field of Psycholinguistics, and who has always made me feel extremely welcome in the lab.

Coming from a Computer Science background, I was initially afraid I would not have the necessary knowledge to pursue this PhD. I was lucky however to have found people along the way in the lab that helped me fill a lot of my knowledge gaps. In particular, I would like to thank Philipp Blandfort, with whom I had great conversations about AI, semantics, exams, students and courses, and who, despite having interacted only for a short time with me, helped me a lot to find my way in my new area. I also would like to thank every one of the many interns we had over the years (mostly from Northeastern University, in the US), all of whom were pretty helpful, not only to my work (where they helped with so many things, such as grading exams, making exam questions, annotating data, programming, running online questionnaires, finding papers, ...), but also to my professional development (as a [Psycho]Linguistics researcher, as a "manager" [of people], as a teacher, ...) as well as to my personal development. Thus, I would like to thank Jasmine Segarra, Abigail Hodge, Mark Murphy, Luc Henriquez, Jialin Selena Song, Hannah Lee, Anja Castro-Diephouse, Rhiannon Stewart, Adrean Valverde, Lydia Bell, Ryan Paulů, Jade Anglin, and Marion Anglin; and to extend a special thanks to Leah Doroski (who inadvertently showed me that, yes, I had made the right decision in moving to Linguistics – and also accidentally made me start taking care of the plants in the lab), Mateo Vargas-Nuñez (with whom I had long conversations about flags, race, and Latin America), and Shaiban Alshaibani (from whom I accidentally learned a new vowel contrast, and whom I am happy to call a friend nowadays). Still related to the lab, I also would like to thank "the Mensa group" (or "the lunch bunch", as Leigh likes to call it), a set of friends with whom I have had, for years, over lunch, some of the most entertaining (but also useful!) conversations: Boran Alamro, Christopher Allison, Franzi Hekele, Omar Jubran, Fenia Karkaletsou, Sofia Linsenmeyer, Laís Muntini, Sophie Thommes and Maximilian Wolkersdorfer.

Indeed, "friendship" was one important theme in my life during these years, and, truth be said, it was the presence of friends (even if sometimes only virtual) that made the path a bit easier in moments of hardship. The word *thanks* may generally sound too little to express the gratitude I feel and the value I place on such friendships, but it is precisely through the quirks of friendship that I know they will understand its actual meaning. Thus, I would like to say *thanks* not only to the friends I have already mentioned earlier, but also to Wesley Mateus

Becker, Christopher Andrews, and Suvrath Padmanabha Pai, who, despite not having been necessarily around through the whole PhD, were instrumental in various moments; to some of my "virtual friends" Putu Agus Khorisantono, Marcus Vinicius Santos, and Ahmed Hariz, who, in various moments, accompanied everything going on despite not being physically there; to my "undergrad friends" Luca Couto Manique Barreto, André Antunes da Cunha, Kim Aragon Escobar, Bruna D'Andrea de Andrades, Fábio da Fontoura Beltrão, Fabiana Bender, Mônica Rousselet Farias, Aline Flor, Lucas dos Santos Lersch, Éderson Vargas Vieira, Larissa Emy Ushiwata, and Lucas Fialho Zawacki who have been especially important, and without whom I would neither have been able to get through the pandemic, nor have reached the point of writing these words right now; and to the more "loose connections" Benhur Golowniczy Brião, José Luís Damaren Júnior, Eliasibe Luis de Souza, Ruohan Gao, Pramod Guruprasad, Hugo Robalino, Mariia Naumovets and Francisco Rocabado, who every now and then spawn back into existence in my life, and whom I do not think I will ever completely lose contact with anymore.

Finally, there is family. This thesis is dedicated to three people who made much sacrifice in order to allow me the opportunity of coming to study (and eventually live) in a different country. In addition to them, I thank my cousin Jussara Borges for always offering me academic support (and just being a great cousin in general); my siblings Jéssica Longaray Gambôa, Thiago Longaray Gambôa, and Suellem Longaray Gambôa (who I hope one day will actually read this here), and especially my brother Jean Daniel Borges Gambôa (along with his family), whom I will not thank for anything in particular lest it become a joke-comment in family gatherings for the rest of our lives.

Abstract

Scientific texts are commonly known to be hard to read. This may be for a number of reasons, such as the fact that they contain a large number of nominalizations, they often use complicated jargon, and sometimes even assume reader knowledge. In order to find ways to make scientific texts clearer, it may be worth to better understand the characteristics of this register. In this thesis, I focus on one aspect of scientific texts: their frequent use of a structure I refer to as complex nominal compounds.

Nominal compounds (structures such as *research paper*, *linguistic analysis paper*) are composed of a head noun (*paper*) and one or more modifiers (*research*, *linguistic*, *analysis*), which can be either nouns or adjectives. *Complex* nominal compounds are nominal compounds made up of three or more words. Despite decades of research on nominal compounds (complex and otherwise), little is known about how complex nominal compounds are processed, or about how they are used in scientific papers. In this thesis, I take steps towards filling this gap.

In my attempt to do so, I recruit two frameworks of language processing and use, through which I produce predictions for the experiments I report here: the Entropy Rate Constancy (ERC) Principle and the Uniform Information Density (UID) Hypothesis. Of particular relevance, not much attention has been given to the UID Hypothesis from the point of view of comprehension. These experiments, therefore, not only contribute towards the understanding of nominal compounds, but also test the validity of these frameworks in general, and of the UID Hypothesis from the perspective of comprehension in particular.

On the processing front, the thesis analyzes the L1 and L2 processing of complex nominal compounds, comparing them with a different structure that was predicted to be easier to process: nouns followed by prepositional phrases (e.g., *paper on the analysis of language*). The results do show that compounds are harder to process than the alternative structure, but were not as clear as predicted. On the usage front, the thesis analyzes the distribution of compounds in a corpus of 182 scientific papers in the fields of Biology, Linguistics and Economics. Compounds appear with roughly the same frequency throughout the different regions of the papers (i.e., do not cluster in specific areas), and are not reused much after the first use. They are also typically set up by their context, and this does have an impact on the difficulty experienced by readers when encountering them (at least when encountering unfamiliar compounds).

The results corroborate the recommendations from writing guides suggesting that compounds should be used with parsimony, but also suggests that familiar compounds do not need to be avoided as much, and that providing contextual support may mitigate some of the difficulties experienced by readers when encountering these structures. It is my hope that future writing guides will take these findings into consideration.

In addition, the results partially support the ERC Principle and the UID Hypothesis, but were not as clear as predicted, and raise questions about the validity of the ERC Principle in texts and of the UID Hypothesis for comprehension.

1 Introduction

Contents

1.1 Motivation **1**
1.2 Structure **5**

1.1 Motivation

> The classic manuals, written by starchy Englishmen and rock-ribbed Yankees, try to take all the fun out of writing, grimly adjuring the writer to avoid offbeat words, figures of speech, and playful alliteration. A famous piece of advice from this school crosses the line from the grim to the infanticidal: "Whenever you feel an impulse to perpetrate a piece of exceptionally fine writing, obey it – wholeheartedly – and delete it before sending your manuscript to press. *Murder your darlings*."
>
> PINKER, 2014, PAGE 12

As I started writing this thesis, I looked up the autocompletions offered by Google for the search prompt "why are scientific papers ...".[1] I wanted to argue that the scientific register in English is complicated to understand and supposed that one of the suggested completions would be along the lines of "... hard to read". Given that my single response would probably be biased by my search history, I also decided to ask friends to send me a screenshot of the autocompletions they were given for the same search prompt. As a result, I received answers from 18 friends who were at the time in Brazil, Canada, Germany, Poland, Sweden and the US. Figure 1.1 lists the main autocompletions we saw. The results did not disappoint:[2] even Google's search system seems to have internalized the idea that scientific papers have their own, difficult style.

Of course, this was not a rigorous experiment. Still, even among academics there has long been criticism about the complexity of scientific texts. The criticism takes form in various ways. An article may claim scientific articles contain "dense, uninspiring language that can be laborious to wade through and difficult to understand" (Doubleday & Connell, 2017, p. 1); another may denounce "the violence" scholars "commit against the English language" (Limerick, 1998, p. 199).[3] The very first lines of one book on "Stylish Academic Writing" states: "For many academics, "stylish academic writing" is at best an oxymoron and at worst a risky business" (Sword, 2012, p. vii). As a result, we are left with "'how to read publication guides'[, which] brace the reader for the hard work ahead" (Doubleday & Connell, 2017, p. 1).

Indeed, scientific articles have become *worse* in the last few decades. For example, the number of syllables of each word, the length of the sentences, the number of "difficult words",

[1] For the sake of clarity and accuracy, I should note that the prompt was actually "why are scientific papers", without the "...". If I added the "...", I received different results.

[2] The results were surprisingly consistent: the four most frequent autocompletions appeared to all but one of the people. Interestingly, one friend sent me Bing results, and they did not include any autocompletion that suggested that scientific articles are hard to read (although my own attempt did include "are so hard to read" as the very first result).

[3] The original sentence was "The politically correct and the politically incorrect come together in the violence they commit against the English language." Limerick's text was especially focused on the different ideological streams that are common in University environments, but these are not relevant to this thesis.

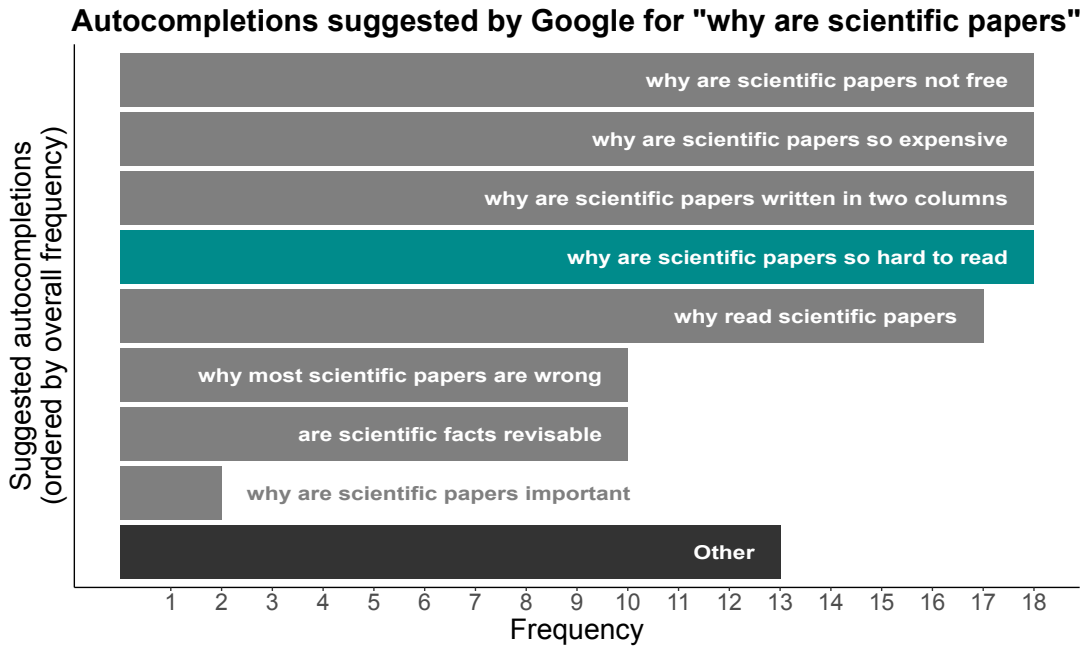


Figure 1.1: Number of times each autocompletion appeared for me or my friends (N=19). Autocompletions with the same frequency are ordered by average rank.

and the number of acronyms used in paper abstracts has steadily increased in the last century (see Plavén-Sigray, Matheson, Schiffler, & Thompson, 2017; Barnett & Doubleday, 2020). As Plavén-Sigray et al. showed, at least some of these measures do reflect trends that extend to other parts of the articles.[4] This kind of writing, despite being quite formulaic in nature,[5] is also inaccessible to the general public, potentially limiting the reach of the research: "Increasing physical access to science (e.g., open access) ... will have less effect if the writing is impenetrable" (Doubleday & Connell, 2017, p. 1). Indeed, we nowadays find movements of researchers producing "Plain Language Summaries" of their findings, so that the non-expert audiences can accurately understand the impact of their research on their lives (see, e.g., Stricker, Chasiotis, Kerwer, & Günther, 2020).

In order to improve this situation, it would make sense to try to understand the characteristics of the register, and how they affect the difficulty people perceive when reading scientific texts. The register has a number of characteristic features, many of which are the focus of several of the writing guides I have cited so far. Sword (2012), for example, devotes a whole chapter to the avoidance academics seem to have to referring to themselves in the first person (*I*), sometimes preferring a royal *we*, sometimes referring to themselves in the third person (*this author*), sometimes using impersonal pronouns (*one*), or even eliminating the agent altogether by structuring their sentences in an agentless passive voice (*as will be demonstrated in this paper* – who will demonstrate?). One other commonly mentioned characteristic of this register (more relevant to this thesis), is its frequent use of nouns. Verbs such as *distinguish*, *examine* or

[4] Interestingly, articles with less readable abstracts (Gazni, 2011) or with less readable language generally (Stremersch, Verniers, & Verhoef, 2007; Hartley, Sotto, & Pennebaker, 2002) seem to be cited more, and Hartley et al. (2002) notes that "there is some evidence that journals do become less readable as they become more prestigious". This contrasts with other research showing that, at least in the Climate Change Science, articles that follow a "Narrative style" receive more citations, and that higher impact journals more frequently follow such a narrative style (Hillier, Kelly, & Klinger, 2016).

[5] Indeed, the strong formulaicity of this register has led many scientific writing guides to contain lists of words and expressions that are commonly used in certain situations (e.g., Bailey, 2011; Hrdina & Hrdina, 2009; Glasman-Deal, 2010; Katz, 2009), or that are typically misused (e.g., Katz, 2009). Hrdina and Hrdina (2009) is a whole book, focused on German L1 speakers, whose sole purpose is to list and categorize many of these expressions.

investigate are often converted into their noun counterparts (*distinction*, *examination*, *investigation*)[6], and sometimes expressions such as *this book originated in* or *he is entirely in favour of* are subsequently referred to with nouns such as *the book's genesis* or *his sympathy for* (examples from Siepmann, Gallagher, Hannay, & Mackenzie, 2011). These conversions, typically referred to as nominalizations, serve a number of purposes. Kerz (2004) argues, for example, that they render the descriptions (of experiments, or results) independent of the experimenters, "not tied to specific conditions or specific observers" – making the text sound "static, rather than dynamic in nature" (p. 126) and the arguments sound "not accessible to debate" (p. 127). That is, by using expressions such as *an analysis of X yielded* (as opposed to *we analyzed X and concluded ...*), authors can abstract away from details of the experiment (including their own agency in conducting it). In addition, Kerz argues that nominalizations are versatile "condensation devices". For instance, she offers Example 1 below, noting that, through the process of nominalization (resulting in Example 1b), what used to be a full sentence (i.e., Example 1a – extracted from the British National Corpus; Davies, 2004) can be flexibly used as a constituent of another sentence (as in Example 1c). In particular, by nominalizing *analysed* into *analysis*, and transforming the sentence object (*data from neurobiologically ...*) into a prepositional phrase (*of data from neurobiologically ...*), it is possible to transform the sentence in Example 1a into the noun phrase in Example 1c.

(1) a. *Lishman and McMeekan (1977) analysed dichotic listening data from neurobiologically intact subjects in terms of familial sinistrality* [...]
b. *Lishman and McMeekan's (1977) analysis of dichotic listening data from neurobiologically intact subjects in terms of familial sinistrality* [...]
c. ***Lishman and McMeekan's (1977) analysis of dichotic listening data from neurobiologically intact subjects in terms of familial sinistrality*** *found* [...], *that among strong left handed individuals* [...]

Given how versatile nouns are, it is not surprising that the noun phrases around them developed into fairly complicated patterns such as the one in Example 1c. Writers attach at will any number of phrases before and after a given (head) noun, specifying very precisely its meaning. Sometimes, however, complicated noun phrases are used to create and refer to new concepts. In these cases, instead of attaching a large number of prepositional phrases to the head noun, it is common for the head noun to be premodified with additional information (Salager, 1984). This gives rise to complex nominal compounds such as *brain tissue autopsy* (i.e., an autopsy of tissue that was extracted from the brain), *waste water treatment facility* (i.e., a facility for the treatment of water from waste), or even *10 megawatt solar thermal electric central receiver Barstow power pilot plant* (the latter example is from Alley, 1996, p. 85). In Bhatia's (1992) words, "the scientific writer frequently needs to refer to very precise and complex concepts again and again and to facilitate that concise reference, he invariably creates compound nominal phrases, which not only promote discourse coherence but also spare him tedious repetitions of long descriptions" (p. 225).

To make this point clearer, consider the following paragraph, from a real scientific article:

(2) *Each session consisted of six trials and all animals completed one session a day. Each session consisted of three correct left and three correct right trials, presented in a pseudorandom order. Each trial comprised two stages, a 'sample run' followed by a 'test run'. At the beginning of each trial, two sucrose pellets were placed in each food well and a metal obstacle was placed at the choice point of the T-maze, thereby closing one perpendicular path (Fig. 8). On a sample run, the animal was placed in the start area and the aluminium obstacle removed, allowing the rat to run down the start arm. Because of the*

[6]Sword (2012) argues that these abstract nouns contribute to the "dull" character of academic texts, devoting another half-chapter to them.

> *metal barrier blocking the entrance to one of the cross arms, the rat could only enter the one open section. Once the rat had collected the sucrose pellets from the well at the end of the open section, the rat was returned to the beginning area, where it remained for 10s while the barrier at the choice point was removed and the same arm as previously visited was rebaited.* (Powell et al., 2017, p. 97)

What would you do, now, if you had to refer to the *barrier* preventing the rats to go back to the *arm* of the T-maze where they *start*ed their trials? This is how the authors of that article continued their paragraph:

(3) *The test run started as the* ***start arm barrier*** *was raised, allowing the animal a free choice between the two cross arms of the T-maze.*

Noun phrases such as *start arm barrier* or *waste water treatment facility*, composed of a head noun and any number of premodifying nouns or adjectives, and typically denoting a single concept, are the main focus of this thesis. In the linguistic and scientific writing literature, they go by many names, such as noun strings (American Psychological Association, 2009), polylogs (Palmer, 1917), pliologs (mistakenly cited by Salager, 1984 when referring to Palmer, 1917), compound nominal phrases (Salager, 1984), complex nominals (Montero, 1996), or run-on noun phrases (Montgomery, 2003) – and many others, see Chapter 2. They have enjoyed a surge of popularity in the last century (Biber & Gray, 2011), having become longer and more complex through time. I will refer to them as *nominal compounds* (NCs).

In particular, in this thesis I will focus on longer, *complex* nominal compounds (CNCs), with three or more words. Despite their popularity, these longer structures are presumed to be hard to process: they have been found to be hard to translate (Carrió Pastor, 2008; Carrió Pastor & Candel Mora, 2013) and hard to paraphrase (Geer, Gleitman, & Gleitman, 1972), and speakers are not able to reliably identify the compound's head (Geer et al., 1972; Limaye & Pompian, 1991). Many writing guides (including some of those referred to in this chapter – Alley, 1996; American Psychological Association, 2009; Siepmann et al., 2011; Montgomery, 2003) even advise against their overuse: "When nouns are packed too close together, like sardines in a tin, the connecting thought gets suffocated" (Tobin, 2002, p. 1534).

Indeed, as I just mentioned, despite their popularity, these constructions are *presumed* (not known) to be hard to process. The studies pointing in this direction have mostly focused on so-called offline behavioral measures (translation, paraphrasing, head identification), that do not collect data while participants are reading, but rather only after they have already read the compounds. Not much is known about the online difficulties experienced by readers (i.e., how the processing happens over time, while comprehending them). In addition, little is known about how complex nominal compounds are distributed in scientific texts or how the context preceding them affects the way they are understood. This leads me to the questions addressed by this thesis. One of the goals of this thesis is to focus on the following two research questions:

Research Question (RQ1) How are CNCs *processed* during reading? In particular, do complex nominal compounds pose difficulties for L1 and L2 sentence processing, as suggested by the literature?

Research Question (RQ2) How are CNCs *used* in scientific articles? In particular, how are they distributed through scientific articles, how are they set up, and how does their set up influence the difficulty perceived by readers when understanding them?

RQ1 is addressed by two papers (see Chapter 5) that compare CNCs such as *waste water treatment facility* with another structure that is also common in scientific papers, but that makes use of prepositional phrases following the noun (e.g., *facility for the treatment of water from*

waste). The first paper (henceforth referred to as **Study 1**) reports on two eye-tracking studies investigating the L1 (English) processing of CNCs, and the second paper (henceforth **Study 2**) reports on two eye-tracking studies investigating the L2 (English) processing of CNCs, with native speakers of German, Portuguese and Spanish. As far as I am aware, this is the first time eye-tracking has been used to investigate the difficulties experienced by readers when processing CNCs.

RQ2 is addressed by two papers (see Chapter 6) that are based on a corpus of scientific articles collected from the fields of Biology, Linguistics and Economics. The first paper (**Study 3**) reports on the distribution of CNCs, and on how they are set up by their contexts. The second paper (**Study 4**) then investigates how this set up influences the difficulty perceived by readers when understanding CNCs.

Answering these questions might help support future attempts to produce recommendations on how to make the scientific register more accessible to the general public. But answering these questions is just one of the two goals of this thesis. In the following, I will first list the two goals, and then explain the relevance of the second goal, and where it came from:

Goal 1 (G1) To answer (or contribute towards answering) the Research Questions 1 and 2.

Goal 2 (G2) To test the validity of the Entropy Rate Constancy Principle for CNC use in scientific texts and the Uniform Information Density Hypothesis for comprehension.

The second goal of this has to do with the way I decided to answer RQ1 and RQ2. Confronted with these questions, one might ask: how *should* CNCs be processed during reading? Should they *really* be difficult? Or how *should* they be distributed, or how much *should* they be supported by their context? In my search for a theory that would propose potential answers for the various aspects associated with these questions, I came across two psycholinguistic frameworks of language processing, both of which are grounded in the ideas of Information Theory (Shannon, 1948): the Entropy Rate Constancy (ERC) Principle, put forth by Genzel and Charniak (2002), and an "extension" of this principle to "all levels of language processing" (Jaeger, 2010, p. 24) known as the Uniform Information Density (UID) Hypothesis. As I will describe in Chapter 3, these frameworks propose that human communication is efficient, that humans avoid transmitting too much information at once, and that, when they do, communication disruptions might occur.

In this thesis, I will use these frameworks to produce predictions for the experiments I report here. Indeed, under the assumption that CNCs are, as I will argue in Section 3.2.2, dense "packages" of information, that convey a lot of information in just a few words, it turns out that CNCs are actually an ideal tool for *testing* these frameworks. Thus, the contributions of Studies 1, 2, 3 and 4 acquire a new dimension: they no longer only investigate the processing and use of CNCs, but rather also aim at evaluating the predictions made by the ERC Principle and the UID Hypothesis. This latter hypothesis, indeed, has been widely investigated from the point of view of production (i.e., on whether the utterances produced by speakers convey information in a uniform way), but not much from the point of view of comprehension (i.e., whether non-uniform speech signal leads to comprehension disruption). This thesis, therefore, contributes to filling this gap, considering the UID Hypothesis from a less well-understood perspective. This is, along with testing the ERC Principle, the second goal of this thesis.

1.2 Structure

The rest of this thesis is structured as follows. In the next chapter (Chapter 2), I introduce in more detail the concept of complex nominal compounds, along with the existing research related to their structure, their processing in L1 and L2, and their use. I proceed with an (extremely

gentle) introduction to the field of Information Theory and to how I intend to use the ideas from this field to investigate the processing and use of CNCs. Using the knowledge from Chapters 2 and 3, in Chapter 4, I pose and justify the predictions associated with RQ1 and RQ2 stated above. Then, in Chapters 5 and 6, the scientific publications that form the core of this thesis are presented, laying out the results of this research, which are briefly discussed in Chapter 7. Finally, Chapter 8 makes a number of concluding remarks and points out directions for future research.

2 Compounding and Nominal Compounds

Contents

2.1 Open and closed classes of words 8
2.2 Creating new words in a language 9
2.3 A note on inconsistent terminology 10
2.4 Noun Compounds 16
2.4.1 What are noun compounds? 16
2.4.2 Structural ambiguity 17
2.4.3 The head noun and its position 18
2.4.4 Relational ambiguity: how are a compound's words linked together? . 19
2.4.5 The scope of this thesis 21
2.5 Previous research on CNC processing 21
2.6 Why are CNCs used in scientific papers? 25
2.7 Summary 26

The YouTube channel *TLDR News*, named after the expression *too long; didn't read*[1], is devoted to publishing relatively short videos covering the main news topics of the moment. Given its success, the channel has expanded into multiple channels, with, among others, a channel focusing on the EU (called *TLDR News EU*), a channel on the US (*TLDR News US* – now defunct), a channel covering other countries (*TLDR News Global*), and a channel with podcasts (*TLDR News Podcasts*). A while ago, they started a new weekly podcast in the TLDR News Podcasts channel, in which they have the faces of many world leaders on a table (a leaderboard), and in every episode each show host moves one leader up and one leader down, depending on how the leaders have fared on the world stage that week. Fittingly, this new podcast is referred to as the *World Leader Leaderboard*.

This chapter is about the phenomena involved in word sequences like *World Leader Leaderboard* or *TLDR News Podcasts*. While these expressions may have seemed easy to parse when I introduced them in the previous paragraph, they actually hide complicated processes that have yielded decades of psycholinguist research. In this chapter, I start by discussing the general processes of word formation, and the general concept of compounding in English. I then funnel into the types of compounds I am interested in in this thesis, first discussing nominal compounds, and then proceeding to the subset of interest, namely complex nominal compounds (CNCs). Finally, I discuss some of the literature involving the processing of CNCs, as well as the reasons why they are used.

[1] The Merriam Webster dictionary defines it as an abbreviation or noun that is "used to say that something would require too much time to read" (Merriam-Webster, n.d.-b).

2.1 Open and closed classes of words

> "What sort of insects do you rejoice in, where *you* come from?" the Gnat inquired.
> "I don't *rejoice* in insects at all," Alice explained, "because I'm rather afraid of them — at least the large kinds. But I can tell you the names of some of them."
> [...]
> "Well, there's the Horse-fly," Alice began, counting off the names on her fingers.
> "All right," said the Gnat: "half way up that bush, you'll see a Rocking-horse-fly, if you look. It's made entirely of wood, and gets about by swinging itself from branch to branch."
> "What does it live on?" Alice asked, with great curiosity.
> "Sap and sawdust," said the Gnat. "Go on with the list."
> Alice looked at the Rocking-horse-fly with great interest, and made up her mind that it must have been just repainted, it looked so bright and sticky; and then she went on.
>
> THROUGH THE LOOKING-GLASS (CARROLL, 1872, P. 55–57)

Before proceeding, it is useful to make a distinction between open and closed classes of words. In everyday life, it is common for speakers to find new concepts that they have not encountered before. In some of these situations, they may decide to create a new word to describe this new concept. Typically, in English, new words are nouns, verbs, adjectives or adverbs. This is because these are English's *open class*es of words, i.e., classes that readily allow for new members[2] For example, the first known use of the word "database" was around 1962 (Merriam-Webster, n.d.-a)[3], when computers were still a new development. As the concept surrounding the word became more popular in the subsequent decades, "database" became a common everyday word in our language use.

In contrast to *open class* words, trying to create new prepositions, determiners or pronouns may encounter resistance from other speakers. This is because they are examples of *closed class*es of words, that do not typically accept new members.

Open class words are the "stuff" of sentences. They convey the main meaning of our sentences, and often have real world referents that we can easily describe. In Example 1, a large portion of the sentence meaning is already expressed only with the bold words, all of which are open class words. This is why open class words are commonly treated as *content words*.

(1) the **rat gnawed** the **clothes** of the **king** of **rome**.

Closed class words, on the other hand, do not normally carry a similarly clear meaning (e.g., what is the meaning of "of"? What about "that"?). Instead, they perform functions, sometimes replacing content words, sometimes indicating whether a content word has already been mentioned before, or sometimes explaining the relationship between other content words. This is why they are often referred to as *function* words.[4] In the same Example 1 above, all non-bold words are function words (in this case, articles and prepositions).

[2] See Fromkin, Rodman, & Hyams, 2011, Chapter 3 for a more detailed discussion on open vs. closed classes of words. Here, we briefly summarize only the points necessary for the discussion.

[3] Sources vary considerably as to its exact first recorded use. While the Collins dictionary (n.d.) suggests it was first used between 1965 and 1970, the Oxford dictionary (n.d.) registers its first use in 1953.

[4] Note, however, that this association between open classes and content words (and its converse, between closed classes and function words) is not as clear cut as it may seem at first. Numerals, for example, are a closed class of words (intuitively, it is not easy to create a new word describing a number), but may arguably be treated as content words in sentences such as "The party received 100 guests".

2.2 Creating new words in a language

> "And there's the Dragon-fly."
> "Look on the branch above your head," said the Gnat, "and there you'll find a Snap-dragon-fly. Its body is made of plum-pudding, its wings of holly-leaves, and its head is a raisin burning in brandy."
> "And what does it live on?" Alice asked, as before.
> "Frumenty and mince-pie," the Gnat replied; "and it makes its nest in a Christmas-box."
>
> THROUGH THE LOOKING-GLASS (CARROLL, 1872, P. 57)

There is a number of ways in which speakers can create new words. For example, they may add an affix (e.g., a suffix or a prefix) to an existing word. The new words in Example 2 are created using this strategy.

(2) a. globe + -al = global (noun → adjective)
b. global + -ize = globalize (adjective → verb)
c. global + -ly = globally (adjective → adverb)
d. globalize + -tion = globalization (verb → noun)
e. re- + paint = repaint (verb → verb)
f. un- + lock = unlock (verb → verb)
g. lock + -ed = locked (verb → verb)
h. lock + -s = locks (verb → verb)
i. car + -s = cars (noun → noun)

As can be seen in the examples, certain affixes can change the class of the original word to which they are attached, so the original noun *globe* gave rise to the adjective *global*, which in turn gave rise to the verb *globalize*, and so on. Still, this is not always the case (*paint* is a verb, and *repaint* is still a verb). In addition, note that the new words in Example 2 are all in open classes. Again, this is not a coincidence: these are the classes that readily accept new members.

There is another strategy that speakers may use in order to create a new word: just juxtaposing them together. This is illustrated in Example 3.

(3) a. oakshield (noun + noun → noun)
b. moonwalk (noun + verb → verb)
c. runtime (verb + noun → noun)
d. blue-collar (adjective + noun → adjective)
e. dark-green (adjective + adjective → adjective)
f. resource-intensive (noun + adjective → adjective)
g. free-ride (adjective + verb → noun)
h. sometimes (pronoun + noun → adverb)
i. without (preposition + preposition → preposition)
j. overrate (preposition + verb → verb)
k. uprise (preposition + verb/noun → noun)
... (see Nakov, 2013, for more examples.)

This juxtaposition of the words[5] into a single unit, quite common in English, is what is typically referred to as *compounding*. The phenomenon is so common that it has been referred

[5]For the sake of accuracy, it is important to note that (as discussed by Lieber & Štekauer, 2011), while introducing compounding as the juxtaposition of "words" works well with English, it would not work well with morphologically richer languages, where the "words" making up the compounds are often required to be in a specific inflection (as is the case in, e.g., Slavic languages), or linking morphemes may need to be introduced in between them (as in, e.g., German). Indeed, Lieber and Štekauer mention that there is great controversy in the literature about what *exactly* the constituents of a compound are: are they "words"? (As suggested

to as "the universally fundamental word formation process" (Libben, 2006, p. 2), "present in all languages of the world (as far as described by grammar)" (Dressler, 2006, p. 23). Dressler called it "the widest-spread morphological technique" (p. 23), stating that any language that can create new words through affixes[6] (as described earlier) must necessarily also have compounding. Of course, many of the words in everyday use are not new. Most of the words in Example 2 and 3 could be easily found in a dictionary, which presumably means they have been in use for a while. However, they are still the result of the word formation process known as compounding.[7]

As can be seen in Example 3, the output of compounding can belong to any open word class (e.g., moonwalk, dark-green) and sometimes even to a closed word class (e.g., without). In this thesis, I am especially interested in the case when the final result is a noun, and in particular in a subset of this group, that I will define in detail in the next sections. Before that, however, I need to address some problems with the terminology.

2.3 A note on inconsistent terminology

It is worth highlighting how inconsistent the literature is in the way it refers to the different constructions that are discussed in this chapter. As briefly pointed out in the Introduction, many names have been used to refer to noun compounds, nominal compounds, and similar structures. This causes much confusion, making, for instance, certain statistics particularly difficult to cite accurately.

Hence, many articles cited throughout this thesis use similar (or identical) names to refer to different things. Table 2.1 shows an overview of the different terms used in the literature, and what they have been used to refer to. To the best of my knowledge, this is the first time such a table has been compiled, even though many authors (e.g., Salager, 1984; Montero, 1996) have long noted the inconsistencies in the literature. Note, however, that it is by no means

by Marchand, 1960.) Or are they "lexemes"? (As suggested by Bauer, 2003.) And how are these specifically defined? A more recent book refers to them as "bases", defining compounds as their combination (Bauer, Lieber, & Plag, 2013, p. 431). While these matters ultimately have implications for the identification of compounds (that is, for deciding whether a certain structure counts as a compound or not), they are out of the scope of this dissertation, and will not be discussed further.

[6] Dressler makes a distinction between two types of affixes, derivation and inflection, that we will ignore here. His exact words were: "if a language has inflection, it also has derivation and compounding, and if a language has derivation, it also has compounding, but not vice-versa" (p. 23).

[7] The description so far, given its informality, begs the question: how to differentiate between a compound such as "greenhouse" and an adjective+noun phrase such as "yellow house"? Several authors (e.g., Levi, 1978, sections 2.3 and 2.4; Bauer et al., 2013, chapters 19 and 20; and Lieber & Štekauer, 2011) have noted how hard it is to define compounds rigorously. In general, a number of criteria have been proposed, that work as "tests" to decide whether a given sequence is a compound or a phrase (e.g., many compounds are written together as a single word – like *greenhouse* –, so, if the sequence is written together then it must be a compound; and many compounds are left-stressed, – like *strawberry* – so this must indicate their compoundness), but these tests either lead to conflicting conclusions (e.g., *floral arrangement* is written separately, but is left-stressed) or simply do not work (e.g., *steel bridge* is written separately, and is right stressed, but would normally be regarded as a compound). A review of the criteria, and of the problems associated with them, can be found in Bauer et al. (2013, chapter 19); but knowledge of these criteria and of their shortcomings will not be necessary for this thesis.

Generally, for my purposes, it will be enough to treat as a compound any noun+noun combination, as well as adjective+noun structures containing so-called "non-predicating adjectives", as described by Levi (1978, chapter 2). These are adjectives that can be used in attributive position (e.g., *a **linguistic** scholar*, *a **chemical** engineer*, *a **provincial** governor* – examples from p. 15), but either do not sound right in predicative position (**a scholar who is linguistic*; **an engineer who is chemical*) or do not keep the same meaning (e.g., *a governor who is provincial* would mean "an unsophisticated governor", and not "a governor of a province"), behaving generally differently from "normal" adjectives (e.g., they normally cannot be modified by *very* – compare *very good engineer* vs. **very chemical engineer*). Levi argued that these adjectives are actually nouns that undergo an optional transformation after being placed in the premodifying position (e.g., *a language scholar* → *a linguistic scholar*). Thus, structures such as *floral arrangement*, *financial benefits* or *dental surgeon* would be compounds, while *convenient arrangement*, *nice benefits* or *wealthy surgeon* would be phrases.

comprehensive: it focuses mostly on papers that have been influential to the research I report here. Indeed, most of the articles included in it are related to compounds made up of three or more words – structures I will refer to as *complex nominal compounds* (see below). It is also particularly arbitrary. No attempt has been made to select the most influential papers, the most cited or the most well-known; no specific journals have been selected or discarded, and no systematic method has been used.

Given the multiplicity of terms shown in Table 2.1, I decided to start this section by explicitly defining what I mean when I say nominal compound, noun compound, and so on. As much as possible, the definitions below will be used consistently in the rest of this chapter and in all parts of the "frame" of this thesis, i.e., outside of the scientific papers that constitute the core of the thesis. In the papers, however, even I have fallen victim to the inconsistencies of the literature: each paper defines its own terms and uses them in a slightly different way.

Compound word A single word, written either "solid"[8] or hyphenated, that is the result of compounding. Examples include nouns such as *snowman* or *blackboard*, but also other word classes (e.g., adjectives: *white-green*, *neverending*; or adverbs: *overnight*, *inside*). Whenever I need to explicitly refer to a compound word of a certain lexical category, I will replace *word* with the lexical category. Thus, *compound noun* is a compound word that is a noun; and *compound adverb* is a compound word that is an adverb.

Noun Compound These are structures that are the result of compounding, and that are nouns. This includes compound nouns such as *snowman* (composed of two nouns) or *blackboard* (composed of an adjective and a noun), as well as sequences of words that, as a whole, constitute nouns (such as *snow ball* and *olive oil*, but also *stomach tissue biopsy* or *waste water treatment facility*) even when not all the words in the sequence are nouns (for example, *floral arrangement*, *financial crisis*, *government economic policy*, *non-invasive medical procedure* or *rail-suspended safety harness*; Williams, 1984). Indeed, as we will see below, even complicated structures such as *"I told you so" face* will be considered noun compounds.

Nominal Compound (NC) Nominal compounds are a subset of noun compounds: every *nominal* compound is a *noun* compound, but not every noun compound is a nominal compound. I am defining the structure much more based on convenience (to explain exactly what kind of structure I have examined in this thesis) and not so much because it is essentially different from *noun* compounds in some particular way.

In order define nominal compounds, it will be useful to make a distinction between the compounding that happens "inside" a word (that produces compound words) and the compounding that links two or more words together.[9] In the case of nominal compounds, I will only be interested in the compounding that links words together, i.e., the compounding of whole words with other whole words. Because of this focus, a nominal compound will need to be made up of at least two words. Compound nouns that are the result of compounding (such as *snowman* or *blackboard*) but that are written solid are not going to be counted as nominal compounds, because they will be treated (from the point of view of this definition) as just a single noun.

I will state the definition, and, in the following paragraphs, try to clarify some important points related to it: a nominal compound is a noun compound that (1) has two or more words and (2) is composed solely of words that, as a whole, are nouns or adjectives (where

[8] I use the word "solid", following other authors (e.g., Bauer et al., 2013; Fernández-Domínguez, 2010), to refer compounds that are written together as a single word (e.g., *moonwalk*). I will refer to compounds separated by a hyphen as "hyphenated" (*dark-green*), and to compounds that are separated by spaces (*olive oil*) as "open".

[9] This distinction is of course arbitrary: some compounds are written sometimes solid, as a compound word (e.g., *healthcare*), and sometimes open (e.g., *health care*). As we will see, the word *healthcare* will not be treated as a nominal compound, but the combination *health care* will.

what constitutes a "word" is decided simply based on spaces). Simple examples include the previously mentioned *snow ball*, *olive oil*, *floral arrangement*, *financial crisis* or *government economic policy*. Importantly, since every nominal compound is also a noun compound, its head word (normally the last word – see Section 2.4.3) is typically a noun, since the head word usually defines the lexical category of the construction as a whole. With this definition out of the way, there are three clarifications that I find necessary to state.

First, note that a nominal compound depends on the definition of "word". In order to count how many "words" a nominal compound has, I will focus exclusively on spaces, and treat hyphenated words and compound words as single words. Thus *non-invasive medical procedure*, *rail-suspended safety harness* and *healthcare policy rearrangement* are all going to be treated as three word nominal compounds (the first one composed of two adjectives and a head noun; the second one composed of one adjective and two nouns, and the third one composed of three nouns).

Second, this definition imposes a restriction on the lexical categories the nominal compound can be made up of: nominal compounds will be composed exclusively of words that (as a whole) are nouns and adjectives. Note that, because nominal compounds are "blind" to the compounding that happens "inside" a word (and only see each word as a whole), I *will* accept words that are the result of compounding with other lexical categories as long as the words, as a whole, are adjectives or nouns. For example, the sequence *neverending task* will be considered a nominal compound, even though it contains the word *neverending*, which includes the adverb *never* in it, but is, as a whole, an adjective: the definition will only care about the lexical category of the word itself, ignoring what it is composed of.

Finally, note that, because nominal compounds only accept nouns and adjectives, it excludes structures such as *"How does it feel?" game*, and *"punishment is good for everyone else, but not my little angel" attitude* (from Goldberg & Shirtz, 2025), that would be considered noun compounds.

N+N Compound These are the subset of nominal compounds that are composed exclusively of words that, as a whole, are nouns. This includes sequences such as *snow ball* and *olive oil*, as well as longer sequences such as *party host disappearance investigation* or *apple pie plate tray accident* (Weiskopf, 2007). As in the case of nominal compounds (NCs), N+N compounds are "blind" to the inner structure of its words. Therefore, a sequence such as *blackboard cleaning material* will be counted as an N+N compound even if *blackboard* is itself a compound word that contains an adjective (*black*).

Complex Nominal Compound (CNC) These are the main focus of this thesis. They are nominal compounds (as defined above) that are composed of 3 or more words. This includes the examples *non-invasive medical procedure*, *rail-suspended safety harness*, *party host disappearance investigation*, and *apple pie plate tray accident*, mentioned above.

With these definitions in mind, it is finally time to describe some of these structures in more detail. In the following section, I focus on noun compounds; but the majority of what is said there is also valid for nominal compounds and N+N compounds.

Table 2.1: Some of the terms used in the literature related to complex nominal compounds. I omitted papers that focused exclusively on the phenomenon of compounding, and that did not refer specifically to nouns that result from it.

Description	Term	Reference	Additional Comment
Compounds composed solely of nouns	Compound noun*	Downing, 1977	
		Horsella & Pérez, 1991	
	N+N combination*	Horsella & Pérez, 1991	The term used was actually *N + N + ... + N combination*
	N+N compound*	Antoniová, 2020	
		Downing, 1977	
		Ferčec, 2015	
		Fernández-Domínguez, 2010	
	NN compound*	Kunter & Plag, 2016	They actually mainly spoke of "NNN compounds"
		Bartolic, 1978	
		Bauer & Tarasova, 2013	
	Nominal compound*	Berg, 2016	
		Horsella & Pérez, 1991	
		Isabelle, 1984	
		Limaye & Pompian, 1991	
		Olshtain, 1981	
	Noun combination*	Kvam, 1990	The paper actually focuses on "three-part noun combinations"
	Noun compound*	Ferčec, 2015	
		Granville Hatcher, 1960	The full term used was "determinative non-appositional noun compounds" (p. 356)
		Olshtain, 1981	
	Noun-noun combination*	Bell, 2012	
	Noun-noun compound*	Cohen & Staub, 2014	
		Gagné & Spalding, 2013	
		Kunter & Plag, 2016	The paper is focused on "triconstituent noun-noun-noun compounds"
		Moroschan, Nicoladis, & Anjomshoae, 2024	

	Noun-noun phrase[1]*	Spalding, Gagné, Mullaly, & Ji, 2010	
	Noun+noun compound*	Bartolic, 1978	
		Downing, 1977	
	Nouns as nominal premodifiers	Biber & Gray, 2011	These papers use miscellaneous explanations saying nouns have been premodified by other nouns
		Dubois, 1982	
		Moroschan et al., 2024	
	Nouns in juxtaposition	Richman, 1969	
Compounds that, as a whole, are nouns	Complex nominal*	Jullian, 2001	
		Montero, 1996	
	Complex nominal phrase*	Bhatia, 1992	Defined them as "series of adjectives, linearly arranged in the pre-modifying position" – but in his example he inadvertedly includes nouns: *Sunpak's advanced video light technology*
	Complex noun phrase*	Carrió Pastor, 2008	
		Carrió Pastor & Candel Mora, 2013	Seem to focus on sequences of nouns, but examples include adjectives
	Compound nominal*	Weiskopf, 2007	
	Compound nominal phrase*	Bhatia, 1992	Defined them as "linearly arranged nouns, occasionally incorporating adjectives as well"
		Salager, 1984	
	Compound noun*	Pérez Ruiz, 2006	
	Compound word*	Geer et al., 1972	
	Conceptual combination[1]*	Gagné & Spalding, 2006	
	Different-component compound*	Ferčec, 2015	
	Modifier-noun combination[1]*	Gagné & Shoben, 1997	
	Nominal Compound*	Berg, 2012	
		Ferčec, 2015	They subcategorize Nominal Compounds into "Different-component compounds" and "Noun compounds" (which they refer to as "N+N compounds" or "N+N expressions"
		Lees, 1960	
		Nagy T. & Vincze, 2013	
		van Helmond & van Vugt, 1985	

	Noun compound*	Nakov, 2013	
		Pérez Ruiz, 2006	
	Noun string*	Pérez Ruiz, 2006	
Single words that are formed through compounding	Compound word*	Dressler, 2006	But some examples include multiple words, e.g., *apple pie*
		Gagné & Spalding, 2006	But some examples include multiple words, e.g., *apple juice seat*, *computer chip*
		Libben, 2006	
Miscellaneous	Nominal compound*	Williams, 1984	Compounds "that start and finish with a nominal, and incorporate an adjective, adverb or participle"
Unclear	Nominal compound*	Marchand, 1955	The title says "nominal compound", but the rest of the paper only uses the term "compound"
		Piera, 1995	Probably meant compounds what I refer to as "noun compounds"

[1] While closely linked and often overlapping to the literature on compounds, the literature on "conceptual combinations" is less focused on the structure (say, whether the modifiers are nouns, or adjectives), and more focused on the so-called "linking relationship" between the modifier and the head noun. See Section 2.4.4.

* Terms marked with a star are terms that, themselves, fit the definition of *nominal compound* used in this thesis.

2.4 Noun Compounds

> "And then there's the Butterfly," Alice went on, after she had taken a good look at the insect with its head on fire, and had thought to herself, "I wonder if that's the reason insects are so fond of flying into candles – because they want to turn into Snap-dragon-flies!"
> "Crawling at your feet," said the Gnat (Alice drew her feet back in some alarm), "you may observe a Bread-and-Butterfly. Its wings are thin slices of Bread-and-butter, its body is a crust, and its head is a lump of sugar."
>
> THROUGH THE LOOKING-GLASS (CARROLL, 1872, P. 58)

Noun compounds make up around 90% of all newly created compounds (Algeo & Algeo, 1991). They are particularly frequent in everyday English use, comprising about 2.6% of the British National Corpus, 3.9% of the Reuters corpus (Baldwin & Tanaka, 2004), and "[i]n the Wiki50 corpus" (Vincze, Nagy T., & Berend, 2011), "67.3% of the sentences on average contain a [noun] compound" (p. 225 Nagy T. & Vincze, 2013).

In this section, I introduce some important characteristics of noun compounds. Since every nominal compound *is a noun compound*, many of the characteristics described here extend to nominal compounds. I start out by describing their structure, then proceed to the distinction between endo- and exocentric compounds, and finally discuss how the words in the compound are linked together. At the end of the section, I then use these ideas to narrowly define the type of structure I am interested in in this thesis.

2.4.1 What are noun compounds?

Noun compounds (e.g., *leaderboard* or *olive oil*) are typically formed by a head noun (e.g., *oil*) and any number of modifying words (e.g., *olive*). As mentioned in the Introduction, English is especially flexible with what can be used in the modifier position. Hence, the modifier can be virtually anything, ranging from a single noun or adjective, all the way to an entire clause (see Examples 4, 5 and 6).

(4) *Adjectives in the modifier position:*

a. blackbird (from Geer et al., 1972)
b. blackboard
c. greenhouse

(5) *Nouns in the modifier position:*

a. toilet paper
b. strawberry
c. blood moon

(6) *Whole phrases in the modifier position:*

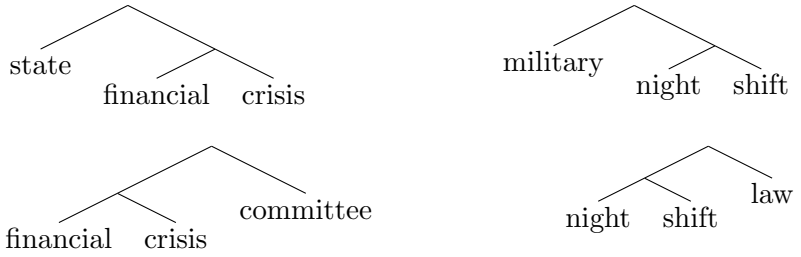
a. floor of a birdcage taste (from Lieber & Štekauer, 2011)
b. wouldn't you like to know sneer (from Lieber & Štekauer, 2011)
c. figure-out-as-you-go process (from Günther, Kotowski, & Plag, 2020)
d. 'Oh , get over yourself' attitudes[10] (from Günther et al., 2020)
e. 'knowledge-of-the-word' test (used in the text of Kvam, 1990)

In most compounds, the modifier extends or constrains the meaning of the head, indicating that they bear *some relationship* with one another (see Section 2.4.4). Thus, *toilet paper* is a specific kind of paper normally used in the toilet, and a *tree branch* is a branch of a tree.

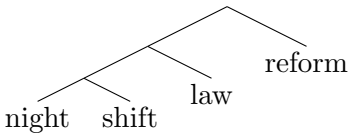
[10]The extra space before the comma is in the original source analysed by Günther et al. (2020).

In some other, more lexicalized cases, noun compounds may acquire specific meanings. This is why *greenhouse*s are normally not green, some birds that are black are not *blackbirds*, and *strawberries* do not really have straws: the combination as a whole denotes a particular concept that is richer than what would be expected by simply putting together its different parts.

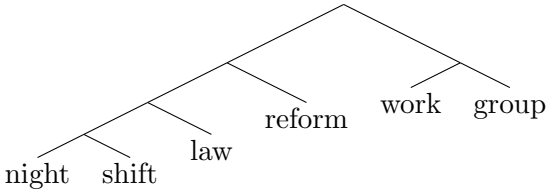
These compounds can then be generally used to form new compounds in a recursive fashion – producing a structure that resembles that of a tree. For example, consider the compounds *financial crisis* (Adjective+Noun) and *night shift* (Noun+Noun). Once they are established as known terms in a certain discourse context, they can be recursively reused to produce new compounds, such as



and of course, the new compounds can be again reused recursively,



with this recursion potentially going on indefinitely:



2.4.2 Structural ambiguity

This recursive structure can, however, be ambiguous sometimes. As noted by Kvam (1990), some three word compounds can form an AB+C structure, some can form an A+BC structure, and still some may be ambiguous in the absence of additional context or world-knowledge. He provided many examples, some of which are listed in Example 7.

(7)

a. **AB+C**	b. **A+BC**	c. **Ambiguous**
birthday party	*concrete lighthouse*	*silk scarf blouse*
shipyard workers	*evening dinner-party*	*brass door knob*
tailbone fracture	*London newspaper*	*silver knife handle*
wartime circular	*nylon clothesline*	*front door step*
soap opera character	*fuel feed pipe*	*stone church tower*
sperm bank donor	*bonnet release handle*	*stell bridge foundation*
health service employee	*parents reply date*	*cotton shirt collar*
income tax relief	*police narcotics control*	*kitchen towel rack*

That is, while it is clear that a *health service employee* is an *employee* of the/a *health service*, and while a it is clear that a *parents reply date* is the *date* by which the *parents* need to *reply*

(to something), it is unclear what a *kitchen towel rack* is: is it a *towel rack* that happens to be in the *kitchen*? Or is it a *rack* for *kitchen towels*, but is free to be located anywhere else? And what about a *silver knife handle*? Is it just the *handle* of a *silver knife*? Or is it the *knife handle* itself that is made out of *silver*?

Note that this ambiguity increases exponentially with the number of words the compound is made up of. While the number of possibilities is quite limited in the case of three words, it grows quickly: a four word compound could have any of the structures listed in Example 8.[11] As we will see later (see Section 3.2.2), this **structural ambiguity** is one reason why, in this thesis, I assume complex nominal compounds (CNCs) to be a difficult structure to process.

(8)
a. A + (B + CD)
b. A + (BC + D)
c. A + (Ambiguous)
d. (A + BC) + D
e. (AB + C) + D
f. (Ambiguous) + D

2.4.3 The head noun and its position

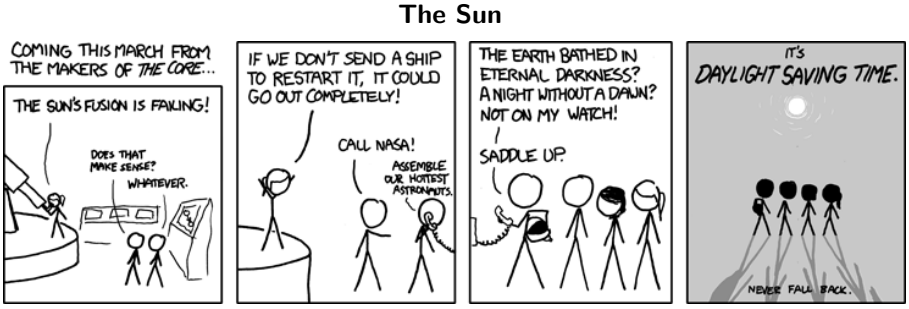
As mentioned in Section 2.4.1, compounds are *typically* formed by a head noun, along with any number of modifiers. Indeed, the majority of the noun compounds given as examples so far have a noun as their last word, and this noun is responsible for a large portion of the meaning of the compound as a whole. That is, most of the time, the compound as a whole is a type/version/instance of its last word: an *adult male rat* is a type of rat, and a *government employee* is a type of employee. This is even the case for many of the lexicalized compounds mentioned so far: a *strawberry* is still a type of berry, and a *greenhouse* could still be interpreted as a type of "house". Similarly, even if there is only one moon, a *full moon* and a *blood moon* could both be treated as different "versions" of the moon (as if they were "different types of moon" that can be observed separately). In all these cases, where the compound has a head noun that lends its own characteristics to the compound as a whole, we say that the compound is **endocentric**.

Note, however, that endocentric compounds may have a head noun that is not its last word. Certain specific cases, such as *attorney general* (a "type" of attorney), have a different structure, where the head is positioned in the left, that is, in the beginning of the compound. Nevertheless, these are still compounds, and still obey the same rules as other endocentric compounds (e.g., they can be used recursively to produce new compounds, e.g., *attorney general senate hearing*).

Endocentric compounds are normally contrasted to other, typically lexicalized compounds, that refer to things that are not a type, version or instance of the head noun. For example, a *honeymoon* is not a type of moon, and a *ladyfinger* is not a type of finger (see Levi, 1978, Chapter 1, for more examples; and p. 6 footnote 3, as well as p. 64 for a discussion of how some exocentric compounds such as *birdbrain* or *cottontail* may have come about through a process known as "beheading"). In these cases, we say that the compound is **exocentric** and that none of the words in the compound are the head noun: the head is outside of the compound itself.

[11] The tree-like structure mentioned here may not always be the best representation for some noun compounds. Nakov (2013) mentions the existence of compounds such as *adult male rat* and *male adult rat*, for which "the order of the modifiers *adult* and *male* could be switched, which suggests that *adult* and *male* independently modify *rat*" (p. 16, emphasis from the original).

2.4.4 Relational ambiguity: how are a compound's words linked together?



Mouseover text: Obligatory bad guy: This operation is sheer foolishness, and it's not happening on my watch! Mainly because I can't figure out how to adjust the time.

MUNROE, 2009

Now consider the meaning of the compound *doll smile*. When read in isolation, readers would normally interpret it as "a smile on a doll" (Gerrig & Bortfeld, 1999). But when read in context, other meanings may be possible. For example, in Example 9, it would be interpreted as "a smile caused by a doll".

(9) Aunt Bev had just come back from visiting her niece and nephew. She was telling her friend Dora about the visit. Bev explained, "I brought Sarah a doll and Paul a baseball."

Dora asked, "Did the children like their gifts?"

Bev replied, "Their mother coached them to give me their biggest smiles." She continued, "I saw a doll smile [...] that was very impressive."

(Gerrig & Bortfeld, 1999, p. 458)

This kind of ambiguity is common in nominal compounds and in compound words. While *baby oil* may refer to a type of oil that is made *for* babies, it might well refer in other situations (maybe a horror movie) to a type of oil that is extracted *from* babies (following the structure of other compounds such as *olive oil*; example from Gagné & Spalding, 2009). Indeed, these are not the only possible interpretations. For example, in a fantasy video-game, there could be an oil that *makes* babies; or if babies were capable of producing oil, this could be an oil *produced by* babies (that is, not extracted in the same way that *olive oil* is extracted from olives).

While these examples might sound exaggerated, the point of this exercise is to show that the words in a compound may be linked together by any number of "relations". This has long been noted in the literature (a review of the ways in which these relations have been considered can be found in Fernández-Domínguez, 2010). While some authors have argued that there should probably be no comprehensive catalog of ways in which the words of the compound can be linked together (e.g., Downing, 1977; Zimmer, 1971), others have tried to produce lists describing all the possible relations between the compounds constituents (e.g., Jespersen, 1942; Granville Hatcher, 1960; Brekle, 1976; Levi, 1978). Interestingly, these lists varied considerably. For example, Granville Hatcher (1960) proposed a total of four very broad relations, that can be seen in Example 10. Thus, the compound *baby oil* would fall into the category A ← B (*baby "is the destination of" oil*), and the compound *olive oil* would fall into the category A → B (*olive "is the source of" oil*).

(10) Linking relations proposed by Granville Hatcher (1960)

a. **A is contained in B**: *gold ring*, *sandpaper*
b. **B is contained in A**: *broomstick*, *seed orange*
c. **A is the source of B** (A → B): *handwriting*, *castor oil*

d. **A is the destination of B** (A ← B): *sugar cane*, *sun worship*

Conversely, Levi (1978) proposed the inventory seen in Table 2.2. She argued that nominal compounds are the result of applying a transformational rule to a phrase containing a noun and a sentence. For example, the compounds *gold ring* and *sports magazine* would originally come from phrases such as *a ring which is made of gold* and *a magazine which is about sports*, respectively. In order to produce the compound, these phrases would undergo a transformation by which the predicate (*is made of gold*, and *is about sports*) is deleted and the noun at the end of the predicate is moved to the modifier position.[12] Thus, the compounds are typically ambiguous: comprehenders need to guess which predicate has been deleted in order to recover the meaning of the compound. This ambiguity, however, is limited to only a small number of (what she termed) **Recoverable Deletable Predicates**, all of which are listed in Table 2.2. Here, I will refer to these as simply "relations". The compound *baby oil*, therefore, would generally fall into Levi's FOR relation; and the compound *olive oil* would fall into Levi's FROM relation. Still, as she notes herself, these relations can be somewhat vague. To see how, she provides the compounds in Example 11, all of which are linked by the FOR relation.

(11) a. *fertility pills* = *pills for fertility* [to increase it]
headache pills = *pills for headache* [to decrease it]

b. *bug spray* = *spray for bugs* [to harm them]
pet spray = *spray for pets* [to help them]

c. *mothballs* = *balls for moths* [of something noxious]
birdballs = *balls for birds* [of something nourishing, like suet]

As she notes, even if *fertility pills* and *headache pills* may be linked by the same relation, they seem to have opposite "goals". "[W]e normally assume that *headache pills* must mean 'pills for suppressing headaches' and *fertility pills* must mean 'pills for enhancing fertility'. My theory claims, however, that all we know for sure is that there is a relationship of intent or purpose between the head noun *pills* and its prenominal modifier (i.e., that the pills are intended to do some **unspecified** thing to, or with, or in connection with, headaches, or fertility ...)" (p. 99, emphasis in the original).

Regardless of the specific way in which the words are linked, and the exact number of relations (four for Granville Hatcher, 1960; twelve in Levi, 1978; and, in a more extreme example, more than a hundred for Brekle, 1976), the main point is that compounds exhibit **relational ambiguity**, and comprehenders often need to choose the correct interpretation on-the-fly, based on world knowledge and contextual information. When compounds are made up of several constituents – as is the case for complex nominal compounds (CNCs) –, this ambiguity is present

[12] For completeness sake, I note that Levi actually proposed two processes for the production of nominal compounds (which she referred to as "complex nominals"): predicate deletion and predicate nominalization. In both processes the compounds are "derived from an underlying NP structure containing a head noun and a full S in either a relative clause or NP complement construction; on the surface, however, the complex nominal is dominated by a node label of N." (p. 50). Predicate deletion is the process I describe in the main text, which has gained the most attention in the compound processing literature.

Predicate nominalization, on the other hand, is the process through which a phrase such as "the act of parents refusing" is turned into "parent[al] refusal". Predicate nominalization is interesting in that it allows compounds to be formed from phrases that are not complete. For example, the compounds "student invention" and "mail sorter" would be formed from phrases such as "x such that students invent x" and "x such that x sorts mail", where x is left unspecified (all examples here are from pages 168, 169, and 173). She categorizes the nominalizations into four types, namely ACT NOMINALIZATION (*musical criticism*, *manager attempts*), PRODUCT NOMINALIZATION (*human error*, *musical critique*), AGENT NOMINALIZATION (*urban planner*, *financial analyst*), and PATIENT NOMINALIZATION (*mammalian secretion*, *presidential appointees*).

Here, I do not discuss this set further, because it has not received much attention in the psycholinguistic literature. A detailed description can be found in (Levi, 1978, chapter 5).

Table 2.2: Each of the twelve Recoverable Deletable Predicates proposed by Levi. More examples can be found in the table in Levi (1978, p. 76), of which this table is a shortened form.

RDP	Example	RDP	Example
CAUSE$_1$	*tear gas*	CAUSE$_2$	*viral infection*
HAVE$_1$	*picture book*	HAVE$_2$	*lemon peel*
MAKE$_1$	*sebaceous gland*	MAKE$_2$	*daisy chains*
USE	*solar generator*	BE	*soldier ant*
IN	*marine life*	FOR	*horse doctor*
FROM	*olive oil*	ABOUT	*abortion vote*

multiple times. As we will see later (in Section 3.2.2), this is one reason why, in this thesis, I assume complex nominal compounds (CNCs) to be a difficult structure to process.

2.4.5 The scope of this thesis

In summary, nominal compounds are a common structure composed of a head noun and one or more modifying adjectives or nouns. In this thesis, I will restrict my focus exclusively to right-headed and endocentric nominal compounds. In addition, I will focus on *complex* nominal compounds (CNCs; which, as defined earlier, are made up of three or more adjectives or nouns) that commonly appear in scientific texts, and are typically not lexicalized.[13] I will investigate the difficulties experienced during the L1 and L2 processing of CNCs. The results reported here are predicted to extend to languages other than English (such as German), but this is not tested in this thesis.

In addition, so far, I have mentioned a number of times that CNCs are a common structure, used frequently in scientific articles. Apart from this fact, however, I am aware of no research that has investigated how they are distributed in these articles. That is, it is unclear whether they cluster in certain parts of the articles, how often they repeat, or whether they are preceded by textual content that is supposed to help with their understanding. If they do receive contextual support, then it is also unclear what strategies are used for introducing them, and how these strategies affect the difficulty perceived by readers upon encountering them. These matters will also be investigated in this thesis.

In the next section, I discuss the existing literature on the processing of CNCs and then proceed to describe some literature on why CNCs are used in the scientific register. I conclude the chapter with a summary of the chapter's main points.

2.5 Previous research on CNC processing

Before discussing the literature on CNCs, a caveat is warranted. The literature on the processing of "non-complex" compounds is vast. In the last few decades, compound nouns formed by the juxtaposition of two words (*warfare*, *greenhouse*) as well as two-word combinations (*coffee pot*, *financial crisis*) have been investigated from a number of different perspectives. Often, they have been lumped together, acknowledging the fact that there are frequent inconsistencies on

[13] In the experiments reported in the publications in Chapters 5 and 6, the items are composed of long sequences of nouns that are not common in dictionaries. However, it *is true* that *parts* of some of those sequences *are* found in dictionaries. For example, the item *United States factory employee insurance costs* contains the two-word noun *United States*, that can certainly be found in most dictionaries. In addition, while the items are not common in dictionaries, at least a few items can be found in Wikipedia. For example, the item *Heart rate variability* does have a corresponding article in Wikipedia (`https://en.wikipedia.org/wiki/Heart_rate_variability`).

whether a combination is written solid, hyphenated or separated (cf. *flowerpot*, *flower-pot* and *flower pot*, Lieber & Štekauer, 2011 – see Kuperman & Bertram, 2012 for a review of the factors influencing these spelling decisions), and the fact that normally the same head noun and modifiers can form both compound words (*lifestyle*, *snowman*) and two-word combinations (*narrative style*, *snow shovel*).

Thus, studies have focused on various themes, such as how compound words are stored in the mental lexicon (e.g., Libben, 2006; Libben, Gagné, & Dressler, 2020), how they are accessed (e.g., Jarema, 2006), how this access is affected by whether compounds are semantically transparent (*sunday*) or opaque (*fleabag*; Marelli, Crepaldi, & Luzzatti, 2009), whether children develop distinct representations for N+N compounds and for their Adj+N counterparts (Moroschan et al., 2024), how speakers process ambiguous compound words such as *clamprod* (that can be interpreted either as *clam+prod* or as *clamp+rod*; Libben, Derwing, & de Almeida, 1999), whether we access "pseudo-morphemes" such as *car* and *pet* or *hip* and *pie* in words such as *carpet* and *hippie* (de Almeida, Dumassais, & Antal, 2020; de Almeida, Antal, & Salehi, 2025), or how we link the two morphemes/words to produce new concepts and whether the linking relations discussed in Section 2.4.4 have any psychological reality (Gagné & Spalding, 2013).

When it comes to CNCs (and similar structures), however, the processing literature is much more limited, and composed exclusively of offline studies. That is, instead of investigating the behavior of participants *during* processing (by measuring, say, reaction times, or eye movements), these studies asked participants to produce a response (say, to respond to a questionnaire, or to paraphrase a compound) and it was these responses that were later analyzed. While many of the results for two-word combinations probably apply to CNCs, few studies have tested the extent to which that is the case. Here, I start by discussing the L1 literature, and then proceed to the studies involving L2 speakers.

CNC processing has been shown to be affected by education level. In one of the earliest studies on the processing of CNCs, Geer et al. (1972) investigated how the processing of structures such as *black bird-house* (and its permutations, e.g., *black-bird house*, *black house-bird*, or *house-bird black*) was affected by the education level of participants, hypothesizing that "[Subjects] from different educational levels differ in grammatical organization" (p. 354). Participants, divided into a higher and a lower education level groups, were asked to (1) produce a paraphrase, (2) evaluate on a four-point scale their degree of confidence about the paraphrase, (3) repeat out loud the compound and (4) again evaluate how confident they were about their recall. Items were either two or three word compounds presented auditorily (which meant that, in order to produce the paraphrases, participants also had to recognize the difference in stress in the pronunciation of the structures – compare ˌ*black* ˈ*bird-house* vs. ˈ*black-bird* ˌ*house*[14]). They found that participants with a higher education level recalled the compounds better and produced more accurate paraphrases. In addition, while the lower education group produced similar paraphrase confidence ratings regardless of their accuracy, the higher education group "evidently had some notion when they were erring" (p. 354), producing paraphrase confidences that did relate to their accuracy. Both groups also differed in their accuracy with two-word compounds (higher education level → higher accuracy), which Geer et al. interpreted as indicative that "the two groups are not in equal possession of all of the relevant rules. They are not equal in competence" (p. 353).

In addition, Limaye and Pompian (1991) have noted that many L1 speakers, when interpreting a CNC in isolation, have difficulties in identifying its head noun. Focusing on long N+N compounds, they made five-item lists of CNCs and asked participants to either paraphrase them or to choose the correct interpretation out of a number of alternatives. Items were presented in isolation (i.e., without any additional context that could help with their understanding) to

[14] Following the International Phonetic Alphabet, I use the symbol "ˈ" to represent primary stress, and "ˌ" to represent secondary stress.

two groups of university students: 75 students of information systems in their "fourth course in data processing and analysis" (p. 11), and 87 students in a business law course, 60% of which were business majors. Each group received a list of CNCs that was related to their course of studies. In their results, they noted that participants, for example, often chose paraphrases such as "management of systems of information" for the compound "management information systems", indicating an inability to identify the CNC's head.

Both aforementioned studies, as well as others (Olshtain, 1981; Williams, 1984) have found that it is not always possible to recover the full meaning of the compounds when they are presented in isolation. Both Olshtain (1981) and Williams (1984) asked participants to paraphrase (C)NCs (and similar structures) and noted that even L1 speakers had low accuracy in the interpretation of certain compounds that required context in order to be understood (e.g., L1 speakers had 33% accuracy on the interpretation of *material defects*, which meant "defects caused by the use of sub-quality materials", because this interpretation is not readily available without context; Williams, 1984).

When it comes to the L2 processing of English CNCs, studies have mentioned the difficulty they cause for speakers of several languages, including French (Williams, 1984), Spanish (Williams, 1984; Carrió Pastor, 2008; Carrió Pastor & Candel Mora, 2013; Jullian, 2001), Croatian (Bartolic, 1978), Hebrew (Olshtain, 1981), Vietnamese and Thai (Linh, 2010). In addition, English L1 speakers learning Spanish have also been found to incorrectly transfer (C)NCs to Spanish, producing structures such as *oro cadena* (literally *gold+chain*), instead of the correct Spanish *cadena de oro* (Richman, 1969). These difficulties have been attributed in the literature to three (interrelated) sources, which are illustrated below: the use of nouns as a premodifiers, word order, and the identification of the head noun.

Regards premodifiers, Bartolic (1978), mentioning the "great difficulties" that English CNCs present "to engineering students whose mother tongue is Croatian" (p. 257), noticed:

> The reason for this lies in the fact that in the Croatian language the nominal structure "adjective + noun" is possible and the "noun + noun" exists only as an appositional structure (rijeka Sava – the river Sava). So only an adjective can modify a noun in Croatian and this structure is an approximate counterpart of the "noun + noun" in English. While this statement generally holds good the factor of usage imposes certain limitations: i.e., the Croatian language has not formed all the possible adjectives which would correspond to the modifying nouns in English. (Bartolic, 1978, p. 257)

Premodifers also present difficulty for speakers of Romance languages such as French, Portuguese and Spanish, where constructions such as *apple tree*, *petrol engine* and *digestion system* have to be translated by the use of single-word translations (Example 12), by prepositions (Example 13), or by converting the premodifying noun into an adjective (cf. *digestive system*; Example 14).

(12) **Apple Tree**

a. **French**: pommier
b. **Portuguese**: macieira
c. **Spanish**: manzano

(13) **Petrol engine**

a. **French**: moteur à essence
b. **Portuguese**: motor a gasolina
c. **Spanish**: motor a/de gasolina

(14) **Digestion/digestive system**

a. **French**: système digestif
b. **Portuguese**: sistema digestivo/digestório
c. **Spanish**: sistema digestivo

Word order differences between L1 and L2 present a second source of difficulty. Romance language speakers are also affected by this second source of difficulty, as illustrated in Examples 13 and 14. While N+N compounds are generally not productive in Romance languages, Adj+N compounds such as *global financial crisis* are typically translated into an "inverted" word order (cf. e.g., Spanish *crisis financiera global*, "crisis financial global"). But that is not always the case. Indeed, when investigating the translation into Spanish of English "premodified complex noun phrases" from scientific texts, Carrió Pastor (2008) found that translators (Spanish L1 medical students with an English B2 level) would produce the most variable words orders. Example 15 shows some examples of the translations they analyzed, with numbers indicating the order in which the words have been translated.

(15) a. *blood urea nitrogen concentrations = concentraciones de nitrógeno en la urea sanguínea*
1 2 3 4 4 3 2 1

b. *more rapidly progressing neuropathy = neuropatía que progresa más rápidamente*
1 2 3 4 4 3 1 2

c. *type-two diabetes mellitus = tipo dos de diabetes melitus*
1 2 3 4 1 2 3 4

d. *PC-SAS version 6.11 = versión 6.11 de PC-SAS*
1 2 3 4 3 4 1 2

Even in languages that do contain N+N compounds, the word order may be a problem. This may additionally lead to the third source of difficulties concerning CNCs in the L2, namely, the misidentification of the head noun. Olshtain (1981) conducted a paraphrase study with Hebrew L1 speakers, a language that forms compounds in an order opposite to English (i.e., instead of *modifier head*, Hebrew compounds follow an order *head modifier*). She found that the Hebrew speaking participants translated compounds such as *laboratory workers* as "a room where they work" or "the laboratory of the workers", indicating a difficulty not only in recognizing the correct order in which the compound should be read, but also in identifying the head of the compound. The same mistake is actually found *in the text* of Linh (2010): "*Voltage source* is a voltage for the source or voltage from the source." (p. 13, emphasis in the original) – just a page after describing this very problem for Thai and Vietnamese speakers of English.

In summary, the studies reported here, despite being exclusively performed offline, indicate that CNCs should be difficult to process not only by L2 speakers, for whom a number of difficulties (associated with their L1) may be present, but also by L1 speakers, who may in fact sometimes not be able to recover the CNC's meaning or even identify its head noun. In Studies 1 (for L1) and 2 (for L2) of this thesis, these difficulties are investigated for the first time using an *online* method, namely, eye-tracking.

2.6 Why are CNCs used in scientific papers?

> "I don't *rejoice* in insects at all," Alice explained, "because I'm rather afraid of them — at least the large kinds. But I can tell you the names of some of them."
> "Of course they answer to their names?" the Gnat remarked carelessly.
> "I never knew them do it."
> "What's the use of their having names," the Gnat said, "if they won't answer to them?"
> "No use to *them*," said Alice; "but it's useful to the people that name them, I suppose. If not, why do things have names at all?"
>
> THROUGH THE LOOKING-GLASS (CARROLL, 1872, P. 55)

In the previous section, I argued that the offline literature on CNC processing suggests that they are probably hard to process. In the next chapter, I will also argue that CNCs are "informationally dense", and that, if that is the case, then they *should* be hard to process. But then, if they cause processing difficulty, *why* do authors keep using them in their texts? Why have CNCs *increased* in popularity in scientific texts over the years (as demonstrated by Biber & Gray, 2011), instead of being disregarded in favor of more understandable alternatives?

A reason may be found in Dubois's (1982)[15] quantitative analysis of the introduction of five experimental articles from the field of Zoology. She initially hypothesized that CNCs had a functional role, i.e., that their use was "determined by the writer's assumptions concerning shared information" (p. 51) with the reader. In other words, CNCs were supposed to signal to the reader that certain pieces of information were assumed by the authors to be already "given", i.e., to be already known. She found that some of the articles did seem to support her hypothesis. In two of them, the introduction was organized in a way that slowly set up the meaning of a very long compound (Example 16 shows the way she schematized this process – Norman, 2003 refers to this process as "packaging"), for example, first using words in postmodifier position (e.g., *the larval stage of D. melanogaster* – where the context clarifies that *D.* stands for *Drosophila*) and only then moving them to a premodifier position (e.g., *Drosophila larvae*). The very long compound then appeared at the very last sentence of the introduction (e.g., after discussing *the regulation of the oxidative NADP-enzymes in larvae*, they eventually arrive at *oxidative NADP-enzyme tissue level regulation*), a "culmination of the process of NP construction" (p. 53). However, for the three other papers she analyzed, the length of the "culminating NP" (this is how she referred to the long CNCs) was pretty modest (e.g., *presumed vestibular function*). Hence, she concluded: "[t]o judge from the five papers being studied, the leftward pileup of modifiers thought to characterize scientific writing is a stylistic feature, i.e., an option, some authors displaying it significantly more than others. Where it is used, however, NP construction serves specific functions" (p. 60).

(16) the NP is X → the X NP is Y → the XY NP is Z ...

Another reason for the emergence of CNCs that has been proposed in the literature (e.g., Levi, 1978, section 3.3.1; Olshtain, 1981; Zimmer, 1971) is the need to refer to new or specific concepts. Downing (1977), for example, mentioned that "compounds often serve as ad-hoc names for entities or categories deemed name-worthy. [...] The more name-worthy the entity or category defined by the compound, the wider the temporal and spatial range of speech situations

[15] Once again, given the terminological inconsistencies of the literature, her work did not focus *exactly* on what I refer to in this thesis as a CNC. In her words, she was interested in the "piling up of modifiers to the left of a head noun" (p. 49) or the "[e]xtensive prenominal modification of head nouns" (p. 64), which, in the terminology of this thesis, fits better what I would refer to as *noun* compounds. For example, in her analysis, she mentions compounds such as *immunologically cross-reactive but enzymatically inactive G-6-PD*, (where *G-6-PD* is treated as a noun). While discussing her work in this section, however, I will still refer to the structures she studied as CNCs. I do this because I am more concerned with the fact that these structures are *complex*, i.e., composed of a head noun and two or more premodifying words.

within which the compound will be useful and interpretable" (p. 841). Salager (1984) discusses this use in the following way:

> There is a semantic difference between the [CNC] banana curve and the related but not synonymous phrase a curve shaped like a banana. In the latter, the curve is identified by the reduced restrictive relative clause shaped like a banana to distinguish it from other curves, whereas the former is a new term in the language for a specific curve having other properties. In other words, the [CNC] is crystallized into a fixed expression owing [*sic*] a scientific meaning which the individual constituents do not have. Therefore, the communicative value of these two expressions is somewhat different. The scientific or technical writer resorts to compounds in the same way and for the same reason as the poet has recourse to metaphors and alliterations. (Salager, 1984, p. 142, emphasis in the original)

This would explain, indeed, why it is in scientific texts that CNCs are so prevalent. In Salager's analysis of 10 texts written in general English (GE) and 10 texts written in a medical English (ME) register, she noted that the "frequency of occurrence [of CNCs] is significantly higher in ME than in GE", and that "the more specialized the text, the longer the [CNCs]" (p. 142).

Finally, a third reason for using CNCs is a need for conciseness, an attempt to save space, an intention to fit a scientific text into a certain number of words or pages (e.g., Levi, 1978, section 3.3.2; Bartolic, 1978; Biber & Gray, 2011; Horsella & Pérez, 1991; Limaye & Pompian, 1991; Olshtain, 1981). Bartolic (1978), for example, discussing the difference between *petrol engine* and *an engine which is driven by petrol*, suggested that the whole material explaining the link between *petrol* and *engine* "is left out for the reason of word economy" (p. 258). He later continues: "[t]his structure is very frequently used in technical writing because it is shorter and more direct and therefore the information is conveyed in a more condensed form which has a greater impact upon the reader" (p. 260). As we will see in the next chapter, this need for "efficiency" may be at odds with the goal of ensuring that the ideas presented in the papers are actually understood by the readers. There might be an optimal level of "compression" of the text that allows for an efficient communication while still guaranteeing that the intended meaning is understood by the reader.

Overall, the literature suggests that CNCs are not just a quirk of scientific texts, but rather seem to offer at least some benefits to their authors. They signal the "given content" to the readers, they act as versatile names that can be reused, and they save space.

2.7 Summary

In summary, compounds are a common structure that is present in every language of the world (Dressler, 2006). In English, noun compounds are the most common form of compound, typically formed with a head noun and any number of premodifying words. From this category, I elected in this thesis a subcategory of interest, *nominal* compounds, that are especially frequent in academic texts (Salager, 1984; Horsella & Pérez, 1991), and have become more and more frequent over the decades (Biber & Gray, 2011). While the processing of two-word nominal compounds has been widely studied, that of longer, *complex* nominal compounds (CNCs) has not. Indeed, little is known about them: it is unclear how they are distributed in scientific texts, how often they are repeated, how the context sets them up, or how this context set up influences the difficulty readers experience when reading them. One of the goals of the research presented in this thesis is to fill this gap.

3 Information

Contents

3.1 A gentle introduction to Information Theory 28
3.1.1 Communication: The typical Information Theory problem 28
3.1.2 A measure of information 31
3.1.3 Entropy 34
3.1.4 Mutual information and the channel capacity 35
3.2 Applying these ideas to Psycholinguistics 37
3.2.1 A possible psycholinguistic instantiation 37
3.2.2 What does this mean for CNCs? 38
3.3 The Entropy Rate Constancy (ERC) Principle 39
3.4 The Uniform Information Density (UID) Hypothesis 41
3.5 Summary 42

Information is a broad concept. The word **information** has many meanings. When paleontologists say that they have no "information" about certain civilizations, when historians consider "information" about a certain historical period, when information scientists investigate the phenomenon of "misinformation", or when computer scientists process "information" contained in a certain database, they all mean and are referring to different things (recovered artifacts, historical records, human communication, and computer data, respectively). And yet, in a way, they are not.

In this chapter, I discuss one specific way in which information has been conceptualized. This is by no means the only one. Just as linguists still today disagree about the definition of "word", or what constitutes the boundary between languages, information scientists and philosophers alike still today debate about the best way to conceptualize information (see e.g., Capurro & Hjørland, 2003; and Floridi, 2009 for discussions on how it has been conceptualized through the years). Here, I put aside these debates and focus on a view that has historically been especially influential for its applications to telecommunications: the view of Information Theory (Shannon, 1948).

In the previous chapter, I discussed the existing literature on complex nominal compound (CNC) processing from a more "traditional" perspective, focusing on the way CNCs are structured, the different ways in which they can be ambiguous, etc. In this chapter, I consider them from the perspective of Information Theory, arguing that they convey a lot of information at once. I will also link them to the two frameworks that guide the predictions I state in Chapter 4: the Entropy Rate Constancy (ERC) Principle (Genzel & Charniak, 2002) and the Uniform Information Density (UID) Hypothesis (Jaeger, 2010).

In order to discuss all of this, I will start with a quick introduction to Information Theory. This introduction will be necessary to explain what I mean with "information", to instantiate the problem to the case of human communication, and to argue why CNCs must be informationally dense. The relationship between CNCs and the ERC Principle and the UID Hypothesis will follow from this description, building upon the concepts introduced in it.

3.1 A gentle introduction to Information Theory

This section is meant as a "quick recap" of the basic concepts of Information Theory. A large portion of the storyline I follow here is inspired by Chapter 8 of Maurits (2012), and by StatQuest with Josh Starmer (2021). This is by no means a comprehensive description: the concepts introduced here are *the bare minimum* I deemed necessary in order to understand the ERC Principle and the UID Hypothesis described later in the chapter. For more information on the topic, the reader is referred to a textbook (e.g., Gallager, 1968).

3.1.1 Communication: The typical Information Theory problem

> When Pfungst (1911) demonstrated that the horses of Elberfeld, who were showing marvelous linguistic and mathematical ability, were merely reacting to movements of the trainer's head, Mr. Krall, (1912), their owner, met the criticism in the most direct manner. He asked the horses whether they could see such small movements and in answer they spelled out an emphatic "NO." Unfortunately, we cannot all be so sure that our questions are understood or obtain such clear answers.
>
> Lashley, 1949, p. 28

As the name of Shannon's paper itself suggests (it is called "A Mathematical Theory of Communication"), Information Theory is a field concerned with communication. The typical scenario is one in which an **information source** is trying to send a message to a **destination** through a **channel of communication**. This message is then encoded into a set of symbols that can be transmitted into the communication channel (by a **transmitter**), and that can be received at the other end (by a **receiver**) and decoded back into the original message. The typical problem in the field is that the communication channel may be noisy, so that some of the symbols input at one end of the channel may end up corrupted when they reach the receiver at the other end of the channel. The goal of the field is to find ways to make the communication robust to this noise, establishing reliable communication, but without requiring too many resources. Figure 3.1 shows a depiction of what this communication framework may look like. In the following paragraphs, I expand on each of the components of Figure 3.1, describing them in more detail.

The information source produces a sequence of symbols $[x_0, x_1, x_2, \dots]$, each of which belongs to a **source alphabet** X (i.e., $x_0, x_1, \dots \in X$).[1] The sequence as a whole represents the real message that needs to be transmitted. For example, the message could be the content of an email, and the symbols could be each character of the email; or the message could be a literal text message from a phone (say, a sequence of words), and the symbols could be the words in that text message. Importantly, these symbols are not completely unstructured: they form the message, and, as we will see in a moment, have some statistical properties (e.g., if the symbols were words, we could know that it is not common in English to finish a sentence with the word "a").

In order to arrive at the destination, the message needs to pass through the communication channel. This channel is described by two components: a **channel alphabet** Y and a probability distribution $P(y_j|y_i)$ that determines the probability that the output of the channel will be the symbol $y_j \in Y$ when the input to the channel was the symbol $y_i \in Y$. For example, if the channel alphabet were composed of the values 0 and 1 (i.e., if the channel values were bits), then one possible channel could be the one represented in the Figure 3.2, where f determines the probability of "flipping" a bit; in other words, of the channel producing as output a 1 when the output was 0, or producing a 0 when the output was 1. When $P(y_i|y_i) = 1$ and $P(y_j|y_i) = 0$ for any choice of different i and j (which would happen, in the example of Figure 3.2, when f is zero), we say that the channel is **noiseless**. In any other case, the channel is **noisy**.

[1] In order to avoid confusion, I will use [and] to represent a sequence, and { and } to represent a set.

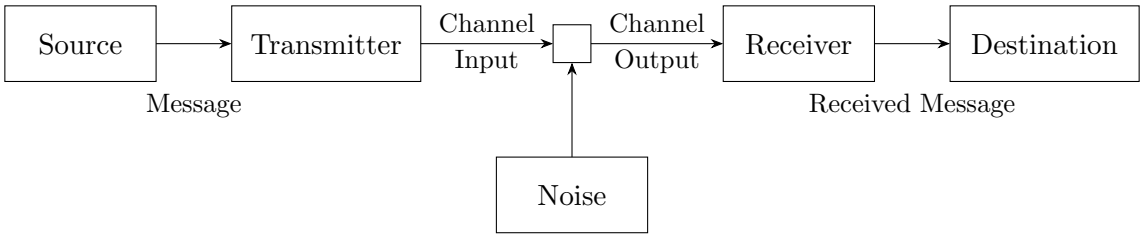


Figure 3.1: A diagram, adapted from Shannon (1948), depicting a "general communication system" described in that paper. The message produced by the information source is encoded by the transmitter into the channel alphabet, and input into the channel, which may be noisy. The output of the channel is then decoded by the receiver. The goal of the field is to achieve reliable communication in an efficient manner.

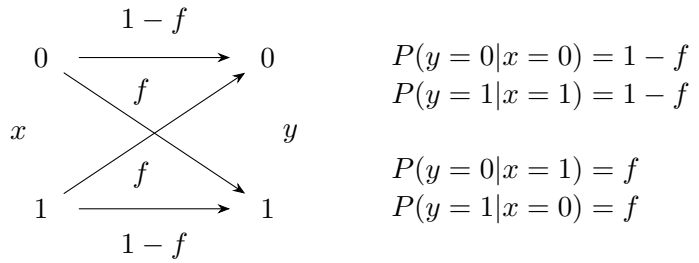


Figure 3.2: An example of a simple communication channel. This channel is a so-called *binary symmetric channel*: it is *binary* in that the input and output are either 0 or 1, and is *symmetric* in that the probability of "flipping" a 0 into a 1 is the same as the probability of "flipping" a 1 into a 0. Adapted from (MacKay, 2003, p. 4)

In order for the message to be input into the communication channel, the symbols $x_i \in X$ (that are part of the source alphabet) need to be converted into the set of symbols $y_i \in Y$ (i.e., into the channel alphabet). This conversion is performed by the transmitter and is referred to as **encoding**. Conversely, when the symbols y_j are received at other end of the communication channel by the receiver, it performs the task of **decoding** the symbols $y_j \in Y$ back into symbols $x_j \in X$.

In a noiseless channel, the output of the channel is guaranteed to be the same as its input. In this case, the goal here is to find a code (i.e., a set of rules for converting each x_i into an y_i) that will minimize the use of the channel, that is, that will lead the messages (i.e., the sequences) $[x_0, x_1, \dots]$ to be converted into sequences $[y_0, y_1, \dots]$ that are, on average, the shortest possible. In this case, the problem of communication boils down to one of **efficiency**. One common example of this kind of communication (cited by Maurits, 2012) could be that of file compression algorithms such as WinZip. Compression algorithms are not really "communicating" anything in the way we are used to thinking of it, but they do offer a good example of how flexible the framework of Figure 3.1 can be. If we try to map each box of Figure 3.1 to a real-world element, the information source would be represented by the file we are trying to compress. This file, composed of a sequence (say, of characters, or of pixels, or bytes) would then be encoded into a compressed form, and input into the channel, represented here by the computer storage device (say, a hard drive, or an SSD). Finally, the "arrival" of the message at the destination would be represented by uncompressing the file, recovering the original file. In this example, the channel is "noiseless" because, once the file has been compressed, we do not expect it to be changed: the channel input (the bytes that are inserted into the computer storage device) is identical to the channel output (the bytes that are read when uncompressing the file).

Now, these file compression algorithms take advantage of the fact that files are not completely random sequences of 0s and 1s, but rather typically contain a lot of repetition and a predictable

Table 3.1: The frequency of each letter in English (according to Lewand, 2000), along with a common binary representation of each of lowercase letters of English. ASCII is the name of the standard used for this binary representation, and stands for *American Standard Code for Information Interchange*.

Letter	**Frequency**	**ASCII sequence**	**Letter**	**Frequency**	**ASCII sequence**
e	0.12702	01100101	m	0.02406	01101101
t	0.09056	01110100	w	0.02360	01110111
a	0.08167	01100001	f	0.02228	01100110
o	0.07507	01101111	g	0.02015	01100111
i	0.06966	01101001	y	0.01974	01111001
n	0.06749	01101110	p	0.01929	01110000
s	0.06327	01110011	b	0.01492	01100010
h	0.06094	01101000	v	0.00978	01110110
r	0.05987	01110010	k	0.00772	01101011
d	0.04253	01100100	j	0.00153	01101010
l	0.04025	01101100	x	0.00150	01111000
c	0.02782	01100011	q	0.00095	01110001
u	0.02758	01110101	z	0.00074	01111010

structure. For example, suppose we wanted to compress a simple .txt file created in Windows's Notepad, containing English text. Our goal is to perform "efficient" communication by "using the channel" as little as possible, i.e., inputting as few as possible symbols into the communication channel. Table 3.1 shows the relative frequency of each English letter. Let us make the simplifying assumption (just for the sake of this example) that the file contains only lowercase letters (and no spaces, punctuation, or capital letters). In that case, the file would contain, for each letter, the binary sequences shown in the ASCII column in Table 3.1, corresponding to the x_i of the file we want to compress. If we wanted to map them into a code y_i (another sequence of bits, in this case, to be stored in the computer), it would be useful to note that, for example, every sequence starts with 011, and so just removing these offending bits (say, replacing 01100101 with 00101) would already "compress" three eights of the file length. But we could go further. Note how, since the text file contains English text, roughly 12% of the 8-bit sequences in the file would be the sequence 01100101 (corresponding to the letter e), about 9% of the 8-bit sequences would be 01110100 (the letter t), and only 0.074% of them would be 01111010 (the letter z). A good compression algorithm (performing efficient communication) would take advantage of these statistics by, for example, mapping the frequent sequences x_i into especially short symbol sequences in the channel alphabet (say, mapping 01100101 into simply 0), and leaving the longer sequences for the infrequent letters.

In a noisy channel, on the other hand, efficiency is not the only variable that needs consideration. This is because the most efficient codes might not be very robust to noise. Say our channel is the one in Figure 3.2 and consider, for example, what would happen if the letter "e" (or the byte 01100101) were encoded into the bit 0, as suggested in the last paragraph. This would indeed achieve very good compression; but then, with probability f, the single bit 0 would be flipped into 1, and what should have been the letter "e" would end up becoming something else. In order to deal with this kind of noise, codes typically introduce some amount of redundancy. One very simple example of what this redundancy might look like (that is, in fact, certainly not very efficient) is what is referred to as *repetition code*, that is, to just repeat a few times every bit being input into the channel. For example, instead of transmitting a single 0 for the letter "e", the transmitter would transmit 00. In this case, if the receiver receives a 01, it knows something went wrong, and can take action to try to fix the communication (say, it can ask the

transmitter to send the message again). Indeed, with more repetitions, it could even *correct* the corrupted message. For example, if the transmitter repeated five times every bit (i.e., sent 00000 instead of 0), and one (or even two) of the bits were flipped (e.g., the receiver received a 01000), then the receiver could still recover the intended message by just performing a "vote" (four votes for 0, and one vote for 1). Ensuring that the messages will be correctly recovered (or recoverable) is what is typically meant with reliable communication.

Of course, repeating each symbol input into the channel several times will cause the channel to be used a lot, reducing efficiency. This is not an accident: the two goals of reliable and efficient communication are generally at odds with one another, and the goal of the field of Information Theory is to find codes that can ensure reliability (i.e., can ensure that the message delivered by the receiver is the same as the message produced in the information source, or, in other words, can ensure that there are no communication errors) while also maintaining as much efficiency as possible. Shannon showed that it is generally possible to keep the probability of communication errors as arbitrarily low as desired as long as the codes being input into the communication channel follow certain properties. In order to discuss these properties, however, we need to discuss the concept of Information.

3.1.2 A measure of information

From the point of view of Information Theory, **information** or **surprisal** is a number that is proportional to how unlikely the outcome of an event is. Here I try to present a friendly justification for the formula used for its calculation. Many of the papers I cite in this thesis that refer to this quantity just state the formula as if it had "come from the heavens", without explaining where it comes from or indicating where I could learn about it. With this explanation, I hope to show that the formula is not as arbitrary as it looks like at a first glance, but rather was chosen because it has certain useful properties that reflect intuitions we have about how "surprising" and "informative" things behave in the real world.

I start with an example,[2] and in this example it makes more sense to focus on the word *surprisal*, instead of information. Imagine I were to repeatedly pick a marble from the following two boxes, A and B, containing red and blue marbles:

- **Box** A: contains 99 red marbles and only a single blue marble;
- **Box** B: contains 50 red marbles and 50 blue marbles.

Now, once I pick a marble from box A, how surprised should I be if I found a red marble? The answer is simple: not at all, since the box A contains many more red marbles than blue marbles. On the other hand, if I had found a *blue* marble, I should be pretty surprised, given how unlikely that is.

Moving on to box B, how surprised should I be if I found a red marble? In this case, both marble colors are equally common, so I should not be surprised finding a red marble, just as I should not be surprised finding a blue marble. Still, whatever color I find from box B, I should probably be a bit more surprised with it than I was when I found a red marble in box A: in box B, I never know what I will find (it may be blue in one pick, then red in another pick, and so on), while in box A I can be pretty certain that I will find a red marble most of the time.

If we were to associate probabilities to these processes, we could define

$$P_A(red) = 0.99 \qquad P_A(blue) = 0.01$$
$$P_B(red) = 0.50 \qquad P_B(blue) = 0.50$$

[2] Here, I partially adapt the very intuitive explanation from StatQuest with Josh Starmer (2021). Still, note that he uses the word *surprise*. I instead use the word *surprisal* because it is the more commonly known term in Psycholinguistics (cf. Levy, 2008).

where the subscripts A and B denote the box the marbles are sampled from.

How could we use these probabilities to define surprisal? Note that, when $P_A(red)$ is very high and I pick a red marble, I am very unsurprised; and when $P_A(blue)$ is very low and I sample a blue marble, I am the most surprised (out of all situations). From this, we can see that there is an inverse relationship between surprisal and probability. If we denote surprisal as I (standing for "information"), and if X is a random variable (that can take the values red and $blue$), then we could tentatively write I as

$$I_i^{tentative}(X) = \frac{1}{P_i(X)}$$

where i indicates the box I am sampling from. (In the following, I will omit the index i when speaking about information generally.)

Now, this formula would have a strange behavior in some situations. To illustrate this strange behavior, consider a third box:

- **Box** C: contains 100 red marbles and no blue marble.

In that case, the probability of picking a red marble would be $P_C(red) = 1$, and $I_C^{tentative}(red)$ would be $\frac{1}{1} = 1$, that is, I would still be "1-surprised" when I found a red marble. This would not make much sense: I would like not to be surprised *at all*, i.e., ideally we should find that the information acquired in this case is 0. So we need a revised formula.

But how should we change the formula? From the description above, we would like a formula that has two desired properties:

1. We would like $I(X)$ to be 0 when the sampling process (the process of picking marbles from a box, in this example) is deterministic;
2. We would like a formula that is inversely proportional to $P(X)$.

Before unveiling the actual formula, I want to add a third item to our desiderata. In this case, I want to appeal to our intuition of information (and not surprisal). Imagine my sampling results were secret, and happened, say, in a secret room, and no one was supposed to know what marbles have been selected. For this example, imagine I am repeatedly sampling only from box A. Imagine a spy manages to peek into the room and see what marble has been sampled every time I pick a marble from the box. Say I pick, in my first sampling "round", a red marble. At this point the spy, looking at my results, acquires $I_A(red)$ information (intuitively, this is true regardless of the exact formula of I_A). Now imagine I pick a second marble, and it is now blue. At this point, the spy acquires an additional amount of information, corresponding to $I_A(blue)$. As I keep sampling, the spy keeps acquiring more and more information about the marbles that have been selected.

At the end of my sampling procedure, how much information has the spy acquired? Intuitively, this should be a sum of the information acquired at each round. This is our third desired property:

3. If the sampling process is repeated multiple times, the final surprisal (resulting from all the repetitions) should be a sum of the surprisals of each repetition.

We can achieve these three desired properties by modifying slightly the current $I^{tentative}(X)$ formula:

$$I(X) = \log\left(\frac{1}{P(X)}\right) \tag{3.1}$$

where the base of the log can be any value, but, for convenience I will assume a base 2, since this is the typical value used in most discussions of Information Theory, and is the value Shannon used in his seminal paper. Indeed, with a base 2, we say that this value is given in *bits*.

Using the log function, we can fulfill the three properties we had envisaged for our notion of surprisal. First, note that $\log 1$ is 0. Thus, when an event is deterministic and we are "not surprised at all", $\log\left(\frac{1}{P(X)}\right) = \log\left(\frac{1}{1}\right) = \log 1 = 0$, i.e., we acquire no information.

Second, note that the $\log\left(\frac{1}{P(X)}\right)$ is still inversely proportional to the probability $P(X)$. For example, as we have just seen, when $P(X)$ is 1, $I(X) = 0$, i.e., no information was acquired. Similarly, when $P(X) = 0.5$, $\log\left(\frac{1}{P(X)}\right) = \log\left(\frac{1}{0.5}\right) = \log 2 = 1$, i.e., one bit of information was gained; and when $P(X) = 0.25$, $\log\left(\frac{1}{P(X)}\right) = \log\left(\frac{1}{0.25}\right) = \log 4 = 2$: two bits of information were gained. As $P(X)$ approaches 0, $I(X)$ approaches infinity, but it is undefined for $P(X) = 0$, which, intuitively, would reflect the idea that it makes little sense to talk about the information acquired from events that cannot happen.

Third, information is additive. To see this, it will be useful to remember that $\log(a \times b) = \log(a) + \log(b)$, and to briefly discuss the probabilities associated with more than one independent event. Back to the spy example, where I picked a red marble followed by a blue marble, what was the probability of having picked them both? Since both events are independent and identically distributed, the probability of the combination of outcomes would be given by multiplying $P_A(red)$ with $P_A(blue)$: $P(red, blue) = P_A(red) \times P_A(blue)$. If we now replace $red, blue$ in the $I(X)$ formula, we get:[3]

$$\begin{aligned} I(red, blue) &= \log\left(\frac{1}{P(red, blue)}\right) \\ &= \log\left(\frac{1}{P(red) \times P(blue)}\right) \\ &= \log\left(\frac{1}{P(red)} \times \frac{1}{P(blue)}\right) \\ &= \log\left(\frac{1}{P(red)}\right) + \log\left(\frac{1}{P(blue)}\right) \end{aligned}$$

In summary, information (or surprisal) is a value that is inversely proportional to the probability of an outcome, that is 0 when the outcome is deterministic, and that is additive when multiple independent events occur (say, in succession). When the base of the log is 2, we say that it is given in *bits*.[4]

Before moving on to discuss how this value is used in Information Theory, I would like to point out that the formula for information is also often written in a slightly different way. In order to arrive at the different form, it will be useful to recall that $\frac{1}{a} = a^{-1}$, as well as the following property of the log function: $\log(a^b) = b \times \log(a)$. With these two ideas in mind, we can see that:

$$\begin{aligned} I(X) &= \log\left(\frac{1}{P(X)}\right) \\ &= \log\left(P(X)^{-1}\right) \\ &= -1 \times \log(P(X)) \\ &= -\log(P(X)) \end{aligned}$$

This last form is more compact, and is just as common in the literature (e.g., A. F. Frank & Jaeger, 2008; Jaeger, 2010; S. L. Frank, Otten, Galli, & Vigliocco, 2015; Schmidtke, Kuperman,

[3] I should note that $I(red, blue)$ is a bit of an abuse of notation. In the next subsection we will see that $I(A; B)$ is a measure referred to as mutual information between the random variables A and B. This is not what is meant here. Here $I(red, blue)$ is just the information acquired by picking both a red and a blue marble from the box.

[4] The word "bit" is a contraction of "binary digit", where "binary" reflects the fact that it can contain only two values, e.g., either *True* or *False*, or either 0 or 1.

Gagné, & Spalding, 2016; Benjamin & Schmidtke, 2023), although many papers (e.g., Levy & Jaeger, 2006; Jaeger, 2010) and textbooks (e.g., Gallager, 1968) also include or even prefer the larger version of the formula.

3.1.3 Entropy

Now I would like to turn my attention to the concept of entropy. In order to understand it, I would like to stick to the story of the spy trying to peek the outcomes of my sampling procedure. Every time a marble is selected from the box, the spy acquires a certain amount of information.

One could ask how much information the spy would acquire for each sample, on average. Clearly, this would depend on the probabilities of selecting these marbles ($P(X = red) = \frac{99}{100}$, and $P(X = blue) = \frac{1}{100}$) and on the amount of information acquired for each possible outcome (selecting a red marble gives us $-\log(\frac{99}{100}) \approx 0.015$ bits of information, and selecting a blue marble gives us $-\log(\frac{1}{100}) \approx 6.649$ bits of information). In general, we can calculate this kind of average by calculating the *expected value* $\mathbb{E}[X]$ of the information. Given a random variable X, the expected value is calculated as:

$$\mathbb{E}[X] = \sum_{x_i \in X} P(x_i) x_i$$

Thus, plugging the amount of information $-\log(P(X))$ into the expected value formula, we can calculate the average amount of information per event:

$$H(X) = \mathbb{E}\left[-\log(P(X))\right] = \sum_{x_i \in X} P(x_i)(-\log(P(x_i))$$

This expected value is typically called *entropy* (typically denoted by the letter H), and is often treated as a measure of how uncertain we are about a given outcome. To see why, I would like to calculate it concretely for the three boxes A, B and C I defined in the previous subsection. This will also hopefully be useful to help clarify the entropy formula.

- **Box** A: contains 99 red marble and only a single blue marble;
- **Box** B: contains 50 red marble and 50 blue marble.
- **Box** C: contains 100 red marble and no blue marbles.

For box A, we have:

$$\begin{aligned} H_A(X) &= \frac{99}{100}\left(-\log\left(\frac{99}{100}\right)\right) + \frac{1}{100}\left(-\log\left(\frac{1}{100}\right)\right) \\ &\approx 0.99 \times (-0.014) + 0.01 \times (-6.644) \\ &\approx 0.081 \end{aligned}$$

For box B, we have:

$$\begin{aligned} H_B(X) &= \frac{50}{100}\left(-\log\left(\frac{50}{100}\right)\right) + \frac{50}{100}\left(-\log\left(\frac{50}{100}\right)\right) \\ &= 0.5 \times (-1) + 0.5 \times (-1) \\ &= 1 \end{aligned}$$

And for box C we have:[5]

$$\begin{aligned} H_C(X) &= \frac{100}{100}\left(-\log\left(\frac{100}{100}\right)\right) + \frac{0}{100}\left(-\log\left(\frac{0}{100}\right)\right) \\ &= 1 \times 0 + 0 \\ &= 0 \end{aligned}$$

What can we learn by looking at these numbers? When I sample from box A, I am typically pretty certain that I will select a red marble: the probability of selecting a red marble is much higher than the probability for the blue marble. But I still have a certain chance of selecting a blue marble. This chance is small, but exists. When I sample from box C, however, this chance does not exist: all marbles are red and there is no way I can select a blue marble at all. Note how the entropies $H_A(X)$ and $H_C(X)$ reflect these properties. The entropy of box C is a flat 0, indicating that I am "0-uncertain" about the outcome: there is no information to be gained about it at every repetition of the sampling process. Conversely, the entropy of box A is a small number. It is still pretty close to 0 (indicating little uncertainty), but it is nevertheless positive.

These two boxes stand in stark contrast with the entropy $H_B(X)$. In this specific case, the marbles have equal probability, and I have no guess as to which color I will select every time I pick a new marble from the box. Every time I do it, I earn, on average, 1 bit of information, corresponding to my "cluelessness" and to the fact that it is either red or blue ("either zero or one"). Indeed, having equal probabilities for every possible outcome (in this case, having probabilities $P(red) = P(blue) = 0.5$), is the only way to achieve maximum entropy. That is, for two outcomes, the maximum possible entropy that can be achieved is exactly 1 bit. For three outcomes, it would be a bit more than one, but still not two:

$$\begin{aligned} H(X) &= \frac{1}{3}\left(-\log\left(\frac{1}{3}\right)\right) + \frac{1}{3}\left(-\log\left(\frac{1}{3}\right)\right) + \frac{1}{3}\left(-\log\left(\frac{1}{3}\right)\right) \\ &= 3\left(\frac{1}{3}\left(-\log\left(\frac{1}{3}\right)\right)\right) \\ &= -\log\left(\frac{1}{3}\right) \\ &\approx 1.585 \end{aligned}$$

And for four outcomes, the maximum possible entropy is 2 bits, intuitively corresponding to the values 00, 01, 10 and 11. The pattern goes on in an intuitive way: 3 bits correspond to equal probabilities with eight possible outcomes, 4 bits correspond to equal probabilities with 16 outcomes, and so on.

3.1.4 Mutual information and the channel capacity

Now, going back to the general communication framework described earlier, consider what happens when the transmitter inputs a symbol y_i into the communication channel. This symbol has a probability $P(y_i)$ of being inserted into the channel, and thus, conveys $I(y_i) = -\log(P(y_i))$ bits of information. Once in the channel, it may be subjected to noise and changed into another symbol, or it may arrive unchanged at the receiving end. Recall that the channel is characterized by the probability distribution $P(y_j|y_i)$ indicating the probability of symbol y_j arriving at the receiver given that the symbol y_i was input. With these ideas in mind, one might ask how much information on average can be acquired about the channel *input* given the noisy output produced by it. This value is called the **mutual information** between the input and the output. In order to keep consistency and avoid a clash in the notation, I will use Y as random

[5]It is true that $\log 0$ is undefined, but it is justifiable to say that $0\log(0) = 0$ because $\lim_{x\to 0} x\log(x) = 0$.

variable denoting the *input* to the channel, and Z as the random variable denoting the *output* of the channel. In that case, the mutual information is given by:

$$I(Y;Z) = H(Y) - H(Y|Z)$$

Before describing the formula in more detail, let us focus for a moment on the term $H(Y|Z)$. I would like to describe what it represents conceptually, without giving too much attention to how $H(Y|Z)$ is calculated.[6] In simple terms, it measures how much we still do not know about Y given that we already know the value of Z. In order to develop an intuition about it, it is useful to consider two cases. First, when Y is identical to Z (which is true when the channel is noiseless, for example), there is no uncertainty left once we observe the value of Y. Therefore, $H(Y|Z) = H(Y|Y) = 0$ (i.e., we already know exactly the input Y once we know the output, which is also Y). Second, when Y and Z are completely unrelated, we learn nothing about Y upon seeing Z, and therefore $H(Y|Z) = H(Y)$. In other words, knowing about Z does not influence at all our uncertainty with respect to Y.

After understanding $H(Y|Z)$ better, the formula for mutual information becomes pretty intuitive: the information we acquire about the input Y when we observe the output Z is the entropy of Y (i.e., $H(Y)$) minus the amount of information we still do not know about Y given our knowledge of Z (i.e., $H(Y|Z)$). Another way of saying this is that the mutual information is the uncertainty of Y minus the uncertainty we still have about Y once we know the value of Z.[7]

Armed with these ideas, I am finally able to introduce the channel capacity. The capacity C of a channel is the maximum amount we can possibly know about the input of the channel after we observe the output of the channel, when considering every possible way in which the input of the channel can be distributed:

$$C = \max_{P(Y)} I(Y;Z)$$

In other words, if I were to keep a record of the mutual information $I(Y;Z)$ between the input Y and the output Z of the channel, and were able to change the way the input Y is distributed however I wanted, then the capacity of the channel would be the maximum value that I would have recorded after having changed the input in every single possible way.

Note that, in a noiseless channel, we know perfectly the input Y when we observe Z. As mentioned earlier, this is a case in which $H(Y|Z) = H(Y|Y) = 0$, and therefore $I(Y;Z) = H(Y) - H(Y|Z) = H(Y) - 0 = H(Y)$. That is, the capacity C of a noiseless communication channel is simply the entropy $H(Y)$.

In order to understand why this value is important, we need to go back to the communication framework of Figure 3.1. Recall that the information source produces symbols x_i, and that these symbols have a statistical structure. Given the probability distribution of these symbols, we could calculate the entropy of the information source H_S. This entropy indicates how much information, on average, each symbol produced by the information source conveys. For simplicity's sake, let us assume that there is a one-to-one correspondence between each symbol x_i produced by the information source and each symbol y_i, into which x_i is encoded in order

[6] The value $H(Y|Z)$ is the **conditional entropy** of the distribution of Y given that Z is known. It is given by

$$H(Y|Z) = \mathbb{E}[H(Y|Z = z_i)] = \sum_{z_i \in Z} P(z_i) \sum_{y_i \in Y} P(y_i|z_i)(-\log(P(y_i|z_i))).$$

[7] Interestingly, $I(Y;Z) = I(Z;Y)$, and therefore the mutual information can also be "inverted",

$$I(Y;Z) = H(Z) - H(Z|Y) = H(Y) - H(Y|Z) = I(Z;Y)$$

that is, it is would also be possible to calculate how much we know about the output Z when we know what the input Y is.

to be transmitted (i.e., input) into the communication channel. Let us also assume that the information source produces symbols at exactly the same "speed" as the channel is able to transmit them (we will deal with the more general case in the next paragraph). In this case, if the entropy of the information source is smaller than the capacity of the channel (i.e., if $H_S \leq C$), then the following two things must be true. First, there must be a way for the symbols of the information source to be (encoded and subsequently) transmitted through the channel so that the probability of communication errors (i.e., the probability of the message delivered at the destination being different from the message produced by the source) is as arbitrarily low as one desires. That is, there *has to be a way* for communication to be performed reliably, even if this way is not immediately evident. Second, if the entropy of the information source is higher than the capacity of the channel (i.e., if $H_S > C$), then the converse is true: there is no way for the communication to be performed reliably, and communication errors are guaranteed to eventually happen. Since reliability and efficiency typically at odds with one another, the most efficient communication normally happens exactly when the entropy H_S of the information source exactly matches the capacity C, i.e., $H_S = C$.

However, this is only true if the information source produces symbols at the exact same rate as the channel transmits them. Consider what would happen if the channel were able to transmit symbols at a rate r_{Ch} that is only half the rate r_S of the information source. In that case, there could be a communication "bottleneck" at the channel, and reliable communication would only be possible if the entropy H_S of the source were half the capacity C of the channel. Therefore, if the rates r_S and r_{Ch} differ, then reliable communication is only possible if $\frac{H_S}{r_S} \leq \frac{C}{r_{Ch}}$, and not possible otherwise. The most efficient communication would then normally happen when $\frac{H_S}{r_S} = \frac{C}{r_{Ch}}$.

3.2 Applying these ideas to Psycholinguistics

This whole communication framework is interesting, but how is this all related to Psycholinguistics? What are the "symbols" we are sending when we speak a word or write a sentence? What corresponds to the transmitter? Or to the channel? Do we even need to directly specify such a mapping?

In this section I try to concretely instantiate this framework to the case of human communication. I do this with the aim of explaining the main assumption I make in this thesis, namely, that complex nominal compounds are a particularly informative structure, that conveys a lot of information at once.

3.2.1 A possible psycholinguistic instantiation

Many psycholinguistic papers freely use concepts associated with Information Theory without explicitly specifying this mapping. For example, focusing on "duration, prosodic structure and redundancy in spontaneous speech", Aylett and Turk (2004) suggest that there is an "inverse relationship between language redundancy and duration[, which] improves communication robustness by spreading information more evenly across the speech signal, yielding a smoother signal redundancy profile" (p. 31), but do not directly declare what constitutes the channel or what the information source is. Similarly, Willems, Frank, Nijhof, Hagoort, and van den Bosch (2016) "assume that the language-comprehension system, after processing the first $t-1$ words (i.e., the sequence $w_1, \ldots, w_{t-1}$), is in a state that implicitly assigns a conditional probability $P(w_t|w_1, \ldots, w_{t-1})$ to each potentially upcoming word w_t" (p. 2507) and then calculate surprisal and the entropy based on that conditional probability distribution; but do not clarify that (or whether), say, words are the symbols constituting the channel alphabet. Indeed, in some cases, the application of these concepts is a sort of convenience: a means to an end. For example, because entropy is a measure of uncertainty, it seems to be a good way to quantify

the amount of competition between the relations linking the two morphemes of a conceptual combination (e.g., Benjamin & Schmidtke, 2023).

Other psycholinguistic studies, on the other hand, do try to explicitly map each element of Shannon's communication framework into real world entities. Thus, for example, for Gibson, Bergen, and Piantadosi (2013) and Gibson et al. (2019) the source and destination are people, the source alphabet is a set of meanings, which are encoded into utterances and sent through the channel, in turn comprising the "acoustic" or the "visual environments" (Gibson et al., 2019, p. 3); and for Maurits (2012) "the source and destination are the minds, ... the source alphabet is the set of ideas which can be expressed by [them, t]he channel is comprised of the sensory-motor systems ... of the communication parties and also the physical medium of the atmosphere"(p. 128), and the channel alphabet is comprised of phonemes.

As we will see later in this Chapter, the theories I focus on in this thesis are less like the latter studies and more like the former ones. Instead of explicitly mapping the real world entities or describing the channel in more detail, they take an approach similar to Willems et al. (2016) and just focus on the probability of occurence of each word in order to calculate the information conveyed by it. I judged it useful, however, to present in the next paragraph a tentative mapping of the entities of Figure 3.1 into the real world. As I already mentioned, I do it in order to justify in a more intuitive manner the main assumption of this thesis: that CNCs are informationally dense, i.e., that they convey a lot of information in just a few symbols. Still, I make no strong commitment to this mapping, since it is not necessary for the predictions I will state in Chapter 4.

Thus, similarly to Maurits (2012), I treat the source and destination as the minds of human beings, and the source alphabet as the thoughts or ideas that are produced by these minds. These thoughts are encoded into a sequence of written words (in my case, I am exclusively interested in communication that is performed through text, since I am focusing on the academic/scientific register), which are elements of the channel alphabet. The channel itself, thus, is the physical paper on which they are written. The goal of the decoder is to make sense of the written words and deliver back into the reader's mind the meaning that was intended by the text author. Therefore, the noise of the channel comes from at least the following three sources. First, many words are ambiguous, and the reader has to decide in what sense they have been used when decyphering the meaning intended by the authors. Second, the ways in which these words are related is often also ambiguous. Third, some words, especially those involved in scientific jargon or in the jargon of a certain field, may not be recognized by the reader. This is especially problematic for L2 speakers, who may even in fact not recognize some words that L1 speakers would have no trouble dealing with.

3.2.2 What does this mean for CNCs?

With this mapping in mind, I can finally justify why, in this thesis, I treat CNCs as structures that are informationally dense. Indeed, it is interesting to note that other authors have spoken about these structures as having "high information content" (Moon, 1997, p. 56), as conveying "highly compact information" (Linh, 2010, p. 5), having "the capacity to condense large amounts of information in few words" (Jullian, 2001, p. 239), and allowing for the "packing and compressing of complex information" (Pueyo, 1996, p. 258); and thus it would seem actually "justifiable" to assume this to be true without any additional argumentation. Still, it is also true that they most likely did not use the word "information" in the sense referred to here.

So why should I assume CNCs to be informationally dense? My reasoning is the following. First, since the CNCs I focus on here are not lexicalized, there is great uncertainty during the reading of the CNC words as the reader moves from one word to the next. In other words, it is hard to predict the next CNC word based on the previous ones. Second, CNCs often *are* the very scientific jargon just alluded to at the end of the previous subsection, containing words that may cause difficulty just by virtue of being infrequent technical terms. Third, CNCs suffer

from a relatively large degree of relational ambiguity: the way in which the words in the CNC are connected is ambiguous, and left for the reader to figure out. For example, as discussed in Section 2.4.4, it is the reader who has to realize, given the context, whether a *doll smile* is a "a smile caused by a doll" or a "a smile on a doll". Of course, the longer the CNC, the higher the degree of relational ambiguity. Fourth, CNCs also suffer from a large degree of structural ambiguity: the reader also has to figure out which words are more closely linked to each other. For example, readers from papers related to Machine Learning have to know that a *recurrent connection weight* is a "weight of a recurrent connection", and not a "connection weight that is recurrent". Again, the longer the CNC the higher its degree of structural ambiguity. Finally, because the head noun is the last word of the CNC, readers often do not know that the CNC is finished until they reach the subsequent word. That is, if the string *recurrent connection weight* were followed by, say, the word *changes*, the reader would have to integrate the meaning of *recurrent connection weight* into the longer CNC *recurrent connection weight changes*, and it would still be unclear (depending on the context) whether *changes* is a plural noun or a third person singular verb. All of these arguments suggest that the probability of the CNC words to occur together is low, and that there is a lot of noise in the process of "decoding" the CNC into the meaning it was intended to convey, i.e., that CNCs convey a lot of information in the few words they are made up of.

In the rest of this chapter, I describe the two theories that guide the predictions I make in the next chapter.

3.3 The Entropy Rate Constancy (ERC) Principle

Garden Path Sentence

Mouseover text: Arboretum Owner Denied Standing in Garden Path Suit on Grounds Grounds Appealing Appealing

Munroe, 2023

Now, given that communication is efficient and reliable when the entropy of the information source is exactly the channel capacity, could it be that *human* communication is efficient and reliable? That is, even if speakers are unaware of the kinds of considerations they make when producing a sentence, could it be that they actually optimize the information they convey so as to also transmit it at a rate close to the channel capacity?

Genzel and Charniak (2002) decided to test this possibility by focusing on communication that is performed through text. Here, I describe their formulation because I will use it when stating the predictions in Chapter 4.

First, define a sequence of random variables $[X_1, X_2, \ldots, X_n]$, representing each of the words of a text, so that X_1 corresponds to the first word of the text w_1, X_2 corresponds to the second word of the text w_2, and so on, up to w_n, i.e., the last word of the text. Let us consider any general word w_i, at any point of the text. If we fix the words $[w_1, w_2 \ldots w_{i-1}]$, then we could define a random variable Y_i with distribution $X_i|X_1 = w_1, X_2 = w_2, \ldots, X_{i-1} = w_{i-1}$. Y_i is the variable that is important to us: if communication is efficient, then it is performed at the channel capacity, and Y_i has always the same entropy, $H(Y_i)$, regardless of the i we choose.

As a second step, let us focus on the sequence of words $[w_1, w_2 \ldots w_{i-1}]$, constituting the

context of w_i. Note that this context can be divided into two parts: a *global* context containing all the words that are not in the same sentence as w_i (i.e., that are in sentences preceding the sentence in which w_i is), and a *local* context, containing all the words in the same sentence as w_i. If we refer to j as the first word in the sentence in which w_i is ($j \leq i$), then we could define the global context as $G_i = w_1, \ldots, w_{j-1}$, and define the local context as $L_i = w_j, \ldots, w_{i-1}$. Of course, if w_i is in the first sentence of the text, then G_i is empty; and if w_i is the first word in its sentence, then L_i is empty. Thus, we can write Y_i as $X_i|G_i, L_i$, and thus $H(Y_i) = H(X_i|G_i, L_i)$. Now if that is the case, then:[8,9]

$$\begin{aligned} H(Y_i) &= H(X_i|G_i, L_i) \\ &= H(X_i|L_i) - I(X_i; G_i|L_i) \end{aligned}$$

Now consider what happens as i is varied from $i = 1$ all the way to end of the text, keeping in mind that $H(Y_i)$ remains constant regardless of the value of i. As we move forward through the text, the global context G_i keeps increasing. However, the local context L_i stays more or less constrained, only increasing until the end of each sentence, and then "resetting" at the start of every new sentence. Therefore, as Genzel and Charniak put it: "Intuitively, we expect the mutual information at, say, word k of each sentence (where L_i has the same size for all i) to increase as the sentence number is increasing. By our hypothesis we then expect $H(X_i|L_i)$ to increase with the sentence number as well." (p. 200). In other words, in order for $H(Y_i)$ to remain constant (which, as discussed in the first step, is what needs to happen in order for communication to be efficient), and given that $I(X_i; G_i|L_i)$ increases as we move forward through the text, we need $H(X_i|L_i)$ to also increase, to "compensate" for the increase in $I(X_i; G_i|L_i)$. That is, if we disregard the global context and only calculate the entropy $H(X_i|L_i)$ (for word w_i), then we should hopefully see a sort of "trend" in which the entropy slowly but surely increases as we move forward through the text.

Their results showed that, indeed, that is the case. They investigated the information in the newspaper articles from the Wall Street Journal available in the Penn Treebank (Marcus, Santorini, & Marcinkiewicz, 1993), and found that, indeed, this entropy did show a positive trend as they went from the beginning to the end of the articles.

What does this mean for CNCs? If the information conveyed by each word increases, and if CNCs are informationally dense, then we should expect at least the following three things. First, we should expect CNCs not to be very common in the introduction of scientific papers, and then to slowly become more and more common as we reach the paper's conclusion. Second, we would also expect CNCs to be slowly introduced through the text, maybe preceded by instances in which their words appear standing alone, or forming simpler structures. Third, once a CNC has been used, we would also expect it to become "part of the context", thus being treated as a "known term" that can be reused frequently without additional contextual support. That is, once used for the first time, CNCs should repeat frequently.

[8] In order to understand this step, it will be useful to refer back to the mutual information formula. In that discussion, I mentioned that $I(A; B) = H(A) - H(A|B)$, i.e., the mutual information between variables A and B is the entropy $H(A)$ minus the uncertainty we still have once we know the value of B. Similar to this formula, we can define the mutual information of two variables, given that a *third* variable is already known. The formula,

$$I(A; B|C) = H(A|C) - H(A|B, C)$$

is pretty much the same as the original one, except that each term is conditioned on the third, known variable. If we isolate $H(A|B, C)$, then we have that $H(A|B, C) = H(A|C) - I(A; B|C)$.

[9] The formula had a typo in the original Genzel and Charniak (2002) publication. It was originally written as

$$\begin{aligned} H(Y_i) &= H(X_i|G_i, L_i) \\ &= H(X_i|L_i) - I(X_i|G_i, L_i) \end{aligned}$$

but it makes little sense to talk about the mutual information of a single random varible. Still, for their purposes, the important point was that one of the terms depended on the global context and the other did not.

These predictions are tested in Study 3 of this thesis. I note that, in the study, they are framed as a test of the UID Hypothesis. This is because, as we will see in the next section, the UID Hypothesis is an "extension" of the ERC Principle. Thus, the implications of that study are just as valid for the UID Hypothesis as they are for the ERC Principle.

3.4 The Uniform Information Density (UID) Hypothesis

1.2 Kilofives

Mouseover text: 'Oh yeah? Give me 50 milliscore reasons why I should stop.'

I DON'T SEE WHY ABRAHAM LINCOLN SHOULD BE THE ONLY ONE WHO GETS TO COME UP WITH WEIRD WAYS TO SAY NORMAL NUMBERS.

Munroe, 2024

Now, if human *textual* communication is indeed reliable and efficient, then could it be that human communication *in general* is so? Indeed, we do not need to restrict this question merely to the choice of upcoming words. Taking into consideration the results from Genzel and Charniak (2002) for text, and those of Aylett and Turk (2004) showing that speakers take longer to pronounce phonemes that are more informative, Jaeger wondered:

> This raises an intriguing possibility. Human language production could be organized to be efficient at *all* levels of linguistic processing in that speakers prefer to trade off redundancy and reduction. Put differently, speakers may be managing the amount of information per amount of linguistic signal (henceforth information density), so as to avoid peaks and troughs in information density. If so, it should be possible to observe effects of this trade-off on speakers' preferences at choice points during utterance planning. (Jaeger, 2010, p. 24, emphasis in the original)

After noting that this could explain why speakers sometimes vary when making choices between *he's* and *he is*, or omit the word *that* in certain clauses (*the person [that] I told you about*), he proposed the Uniform Information Density Hypothesis:[10]

> *Uniform Information Density (UID)*
>
> Within the bounds defined by grammar, speakers prefer utterances that distribute information uniformly across the signal (information density). Where speakers have a choice between several variants to encode their message, they prefer the variant with more uniform information density (ceteris paribus). (Jaeger, 2010, p. 25)

[10] While I frame the Uniform Information Density (UID) Hypothesis as an "extension" of the Entropy Rate Constancy Principle, not many authors have recognized or treated it as such. The only paper I found acknowledging the relationship between the two language frameworks is Xu and Reitter (2018), who wrote: "More recently, the idea of *uniform information density* ... has extended the ERC into a broader framework that governs how people manage the amount of information in language production, from lexical levels to all levels of linguistic representations, e.g., syntax or semantics. [...] Therefore, UID could be viewed as a generalization of the principle of ERC" (p. 148-149, emphasis in the original).

This hypothesis has gathered support in numerous studies demonstrating that, indeed, speakers do seem to optimize the amount of information they transmit through time (see Juzek, 2024, for a critical review, also considering some of the limitations of this research). Taking a step further, one could also ask: if speakers optimize their *production*, could this optimization also extend to their *comprehension*. That is, could it be that comprehenders also expect the information they "receive" to be transmitted at a constant rate? If that is the case, then we should expect comprehenders to have difficulty when receiving too much information at once, since this would presumably surpass the capacity of the communication channel.

This is what Studies 1 and 2 (Chapter 5) focus on. Studies 1 and 2, therefore, investigate the difficulty experienced by comprehenders while reading sentences containing CNCs. They use CNCs to investigate the UID Hypothesis under the assumption that CNCs are informationally dense structures that probably surpass the channel capacity. In addition, this difficulty associated with CNCs should probably be modulated by the context in which the CNC is encountered: if the context is helpful, readers should perceive less difficulty in comprehending them than if the context does not offer any support. This is what Study 4 (Chapter 6) investigates, considering additionally the role of familiarity in the perception of difficulty associated with CNCs. This "UID for comprehension" hypothesis has received less attention in the literature, and Studies 1, 2 and 4 contribute to closing this gap (but see M. X. Collins, 2014; Sikos, Greenberg, Drenhaus, & Crocker, 2017; Meister et al., 2021 for the few papers I am aware of that have considered this perspective).

3.5 Summary

In summary, in the last decades, a number of researchers have tried to model human communication using the framework proposed by Shannon (1948). In this thesis, I assume that CNCs are dense packages of information, which makes them a suitable tool to test two of the ways in which the ideas of Information Theory have been applied to the understanding of human communication: the Entropy Rate Constancy Principle and the Uniform Information Density Hypothesis.

4 Predictions

"*Bulky baggage* can be picked up at the *bulky baggage counter*"

The text on a monitor at the Frankfurt Airport

Contents

4.1 RQ1: predictions for CNC processing 43
4.2 RQ2: predictions for CNC use 45

At this point, it is finally possible to lay out the predictions for the research questions posed in Chapter 1. For the sake of reading convenience, I restate the research questions at the beginning of each section.

4.1 RQ1: predictions for CNC processing

RQ1 How are CNCs *processed* during reading? In particular, do complex nominal compounds pose difficulties for L1 and L2 sentence processing, as suggested by the literature?

As discussed in Chapter 2, the literature on the offline processing of CNCs generally treats them as hard to comprehend, both for L1 readers and for L2 readers. In addition, assuming that compounds are indeed peaks of information density as argued in Section 3.2.2, the Uniform Information Density (UID) Hypothesis would also predict them to cause reading difficulty. Therefore, in the experiments reported in the publications presented in Chapter 5 (Studies 1 and 2), CNCs are predicted to cause reading difficulty to L1 and L2 readers alike.

In those papers, this *difficulty* is measured by means of eye-tracking reading experiments. In order to describe how this difficulty is operationalized in this kind of experiment, a brief explanation about eye-tracking experiments (what data they produce, and how the data are analyzed) will be needed (see Rayner, 1998 for a more detailed explanation of the eye-tracking methodology in general, and Clifton et al., 2016 for a comprehensive review of the kinds of studies in which it has been used).

In a typical eye-tracking study, participants read sentences while having one (or both) of their eyes tracked by a camera (the eye-tracker). The eye-tracker samples the position of their eye(s) at a high frequency rate (for example, in the experiments reported in this thesis, this frequency rate is between 500Hz and 1000Hz) for each of the sentences they read (i.e., each trial). After the experiment, the data is post-processed by the eye-tracking software, and a number of reports are made available to the researchers.

Crucially, while reading, the eye does not move smoothly through the words. Rather, it performs a number of **fixations** (longer stretches of time in which the eye does not move much) and **saccades** (jumps from place to place, in between the fixations). Hence, once the data has been collected, some of the reports the eye-tracking software produces contain information about the fixations and saccades performed for each sentence. These fixations and saccades are what is analyzed.

Measures associated with these fixations and saccades are then assumed to reflect the ease or difficulty participants experience while reading (parts of) the sentences. For example, if participants linger longer at a particular region of the sentence, then that part of the sentence

is assumed to be harder; if participants consistently make saccades back towards the previous words whenever they reach a certain part of the sentence, then those words are also assumed to be harder. Importantly for the kinds of experiments reported in this thesis, these longer reading times (this slow down in the reading of the sentence) are known to "spill over" to subsequent sentence regions: readers also normally have longer reading times at the words that *follow* a difficult region, even if those words are not necessarily difficult to process. How exactly reading times are calculated (whether it is calculated as the length of all fixations around certain words, or whether it is calculated as the length of only the first fixation, etc.) varies from experiment to experiment, and is defined explicitly in the papers of Chapter 5. Thus, to appply these concepts to our concrete case, if CNCs indeed cause reading difficulty, then sentences containing CNCs should lead participants to slow down (i.e., to have longer reading times) not only on the CNC itself, but also in the subsequent text regions, and participants should also make more frequent saccades back towards the CNC.

The expressions "slow down" and "more frequent" in the previous paragraph, however, beg the question: compared to what? In order to decide whether CNCs are indeed "difficult", in the experiments reported in Studies 1 and 2, participants read two types of sentences that were essentially the same in most respects, but differed only in that one of the types contained a CNC, and the other type contained another structure, which we refer to as **Noun followed by prepositional phrases** (NPP), that has semantic content that is very similar to that of the CNC, but "spreads" this content across many more words. These structures are exemplified in Example 1 below. These structures are referred to as the **critical structures**, and the part of the sentence containing it is referred to as the **critical region**. In addition, the experiments reported in Studies 1 and 2 also manipulated the length of the critical structures. The intuition was that longer CNCs would constitute starker "peaks" of information density and thus should lead to more difficulty during reading.

(1) a. **length-3 CNC:** automation legislation advice
b. **length-3 NPP:** advice for the legislation on automation

c. **length-4 CNC:** factory automation legislation advice
d. **length-4 NPP:** advice for the legislation on automation of factories

The idea of the experiments, thus, is the following. If we compare the reading times of the text areas *following* the critical structure (i.e., the text areas following the CNC or the NPP), then participants are predicted to have longer reading times after having read a CNC than after having read an NPP – because the CNC difficulty will "spill over" to those areas as well. Since the only difference between the two types of sentences was the critical structure (in the critical region), any difference in reading times has to be attributed to the structure itself. Similarly, if we calculate the number of times participants "regress" towards the critical region (i.e., how often they move back towards the CNC or the NPP), then participants are predicted to make more regressions to the critical region when it contains a CNC than when it contains an NPP, reflecting the higher difficulty associated with CNCs.

In addition, these reading measures (the reading times of the text areas *following* the critical structure, and the number of "regressions" towards the critical region) will be higher when the critical structure (the CNC or the NPP) is longer, again reflecting more difficulty associated with longer structures.

Finally, a third prediction concerns L2 speakers. L2 speakers may have different expectations, stemming from their L1, on how CNCs are used. Spanish and Portuguese, for example, are languages that do not make extensive use of CNCs, and thus L1 speakers of these languages may assign low probability to these structures. Conversely, German speakers may assign high probability to them, since German is known to make heavy use of compounds. This differential "probability assignment" of L2 speakers is expected to lead to different processing outcomes.

In particular, in Study 2, which investigates the L2 processing of CNCs, we predict L2 speakers whose L1 makes little (or no) use of CNCs (in our case, Spanish and Portuguese) to have more difficulty with them than L2 speakers whose L1 makes frequent use of CNCs (in our case, German). This additional difficulty should result in an interaction between the speakers' L1 and the type of structure being read.

4.2 RQ2: predictions for CNC use

RQ2 How are CNCs *used* in scientific articles? In particular, how are they distributed through scientific articles, how are they set up, and how does their set up influence the difficulty perceived by readers when understanding them?

There are a number of predictions associated with RQ2, which are investigated in Studies 3 and 4. First, as discussed in Section 3.3, according to Genzel and Charniak's (2002) Entropy Rate Constancy (ERC) Principle for texts, the amount of information conveyed by words in texts remains constant through the text, since "the most efficient way to send information through noisy channels is at a constant rate" (p. 199). If that is the case, then, as argued in that section, if we only considered the amount of information conveyed by each word separately (without taking into account the context preceding it), this value should steadily increase as we move forward through the text (because the preceding context should help make some of the words towards the end of the text "expected", reducing their information content). Since scientific papers are "texts", this principle is presumed to apply to them as well. Now, if CNCs are indeed peaks of information density, as argued in Section 3.2.2, then they should be predicted to be relatively uncommon around the beginning of scientific papers, and to increase in frequency as one progresses towards the end of the papers. In Study 3 we operationalize this by dividing scientific papers into three parts, which we refer to as "Introduction" (typically, everything from the beginning of the paper until the point in which the first experiment is introduced), "Middle" (typically, the Methods and Results of the experiment), and "Conclusion" (typically, the Discussion and Conclusion of the paper) sections. Thus, our first prediction is that CNC frequency should be higher in the Conclusion section than in the Middle section, and higher in the Middle section than in the Introduction section.

Second, also following from Genzel and Charniak's principle, CNCs are not predicted to appear "out of the blue". Instead, similar to the compounds studied by Dubois (1982) (see Section 2.6), they are expected to be set up by their context: some of its words are first expected to stand alone, or to appear in postnominal positions in simpler constructions. CNCs are thus expected to emerge only after these (presumably) easier structures have been presented to the readers. This prediction is examined in Study 3.

Third, once used in a scientific paper, CNCs could become established as new terms, that can be subsequently referred to in future parts of the text. Thus, CNCs are predicted, once used, to repeat frequently in a given paper. This follows once again from the ERC Principle, and is also examined in Study 3.

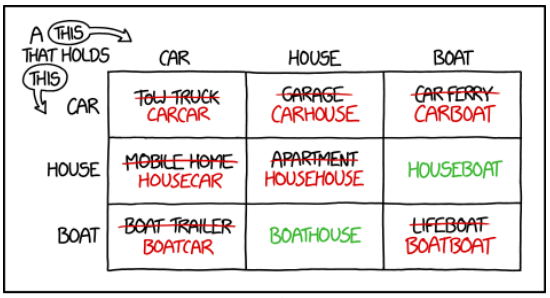
Finally, a fourth prediction, more directly associated with the UID Hypothesis, and not so much with the ERC Principle, concerns what happens as readers go through a text containing CNCs. The amount of support offered by the preceding context to the meaning of a CNC should have an impact on the difficulty experienced by readers when comprehending the CNC. That is, CNCs that receive a great deal of contextual support should be read more easily than those that receive little contextual support. Thus, in Study 4, we operationalize this difficulty by asking participants how much difficulty they experienced, and additionally by asking them to paraphrase the CNCs.[1]

[1]Of course, one does not need a complicated information theoretic framework or the UID Hypothesis to argue

that words that are well introduced by their context are probably easier to understand. Thus, in Study 4, we do not really refer to the UID Hypothesis when stating our predictions. Still, the results of that study does have some implications for how well the UID Hypothesis holds for comprehension.

5 Publications on CNC processing

Boathouses and Houseboats



Mouseover text: The <x> that is held by <y> is also a <y><x>, so if you go to a food truck, the stuff you buy is truck food. A phone that's in your car is a carphone, and a car equipped with a phone is a phonecar. When you play a mobile racing game, you're in your phonecar using your carphone to drive a different phonecar. I'm still not sure about bananaphones.

Munroe, 2018

Contents

5.1 CNC processing in the L1 . . . . . **47**
5.2 CNC processing in the L2 . . . . . **48**

In this chapter, the publications associated with the first research question are reprinted. The publications are inserted as they were published and/or submitted, which means that the page numbers may not align with the page numbers of this dissertation.

The two publications are described in the following sections.

5.1 CNC processing in the L1

The first publication in this chapter is:[1]

Gamboa, J. C. B., Fernandez, L. B., & Allen, S. E. M. (2024). Investigating the Uniform Information Density hypothesis with complex nominal compounds. *Applied Psycholinguistics, 45*(2), 322–367. `https://doi.org/10.1017/S0142716424000092`

The authors contributed to the publication in the following ways:

- **Conceptualization:** Allen, Fernandez;
- **Methodology:** Allen, Fernandez;
- **Software:** Fernandez, Gamboa;
- **Data collection:** Fernandez, Gamboa;

[1] Copyright: Reprinted with permission.

- **Correction algorithm:** Gamboa;
- **Visualization:** Gamboa;
- **Data analysis:** Gamboa;
- **Interpretation of results:** Allen, Fernandez, Gamboa;
- **Writing (first draft):** Gamboa;
- **Writing (review and editing):** Allen, Fernandez, Gamboa;
- **Supervision:** Allen;
- **Funding acquisition:** Allen;
- All authors approved the submitted version of the manuscript.

5.2 CNC processing in the L2

The second publication in this chapter is:

Gamboa, J. C. B., Fernandez, L. B., & Allen, S. E. M. (2025). *Investigating the Uniform Information Density Hypothesis in L2 with complex nominal compounds* [Manuscript submitted for publication]. Psycholinguistics and Language Development Research Group, RPTU University of Kaiserslauter-Landau.

The authors contributed to the publication in the following ways:

- **Conceptualization:** Allen, Fernandez;
- **Methodology:** Allen, Fernandez, Gamboa;
- **Software:** Fernandez, Gamboa;
- **Data collection:** Fernandez, Gamboa;
- **Correction algorithm:** Gamboa;
- **Visualization:** Gamboa;
- **Data analysis:** Gamboa;
- **Interpretation of results:** Allen, Fernandez, Gamboa;
- **Writing (first draft):** Gamboa;
- **Writing (review and editing):** Allen, Fernandez, Gamboa;
- **Supervision:** Allen;
- **Funding acquisition:** Allen;
- All authors approved the submitted version of the manuscript.

Applied Psycholinguistics (2024), **45**, 322–367
doi:10.1017/S0142716424000092

CAMBRIDGE UNIVERSITY PRESS

ORIGINAL ARTICLE

Investigating the Uniform Information Density hypothesis with complex nominal compounds

John C. B. Gamboa, Leigh B. Fernandez and Shanley E. M. Allen

University of Kaiserslautern-Landau, Kaiserslautern, Germany
Corresponding author: John C. B. Gamboa; Email: gamboa@rptu.de

(Received 8 March 2023; revised 19 February 2024; accepted 20 February 2024; first published online 12 April 2024)

Abstract

The Uniform Information Density (UID) hypothesis proposes that speakers communicate by transmitting information close to a constant rate. When choosing between two syntactic variants, it claims that speakers prefer the variant distributing information most evenly, avoiding signal peaks and troughs. If speakers prefer transmitting information uniformly, then comprehenders should also prefer a uniform signal, experiencing difficulty whenever confronted with informational peaks. However, the literature investigating this hypothesis has focused mostly on production, with only a few studies considering comprehension. In this study, we investigate comprehension in two eye-tracking experiments. Participants read sentences of two different lengths, reflecting different degrees of density, containing either a dense structure (a nominal compound, NC) or a structure that spreads the information through more words (a noun followed by a prepositional phrase, PP). Favoring the UID hypothesis, participants gazed longer at text segments following the critical structure when it was an NC than when it was a PP. They also regressed more in sentences containing longer structures. However, the pattern of results was not as clear as expected, potentially reflecting participants' experience with the denser structure or task differences between production and comprehension. These aspects should be taken into account in future research investigating the UID hypothesis for comprehension.

Keywords: Sentence processing; Nominal compounds; Eye tracking

Introduction

A central question in psycholinguistics is how speakers select which variant of a structure to produce when there is more than one option (e.g., brain activation pattern vs. pattern of activation in the brain), and how listeners' comprehension is influenced by which variant is produced. One common way of exploring this question appeals to information theory, originally put forth by Shannon (1948) and developed extensively in psycholinguistic work under the notion of surprisal (Hale, 2001; Levy, 2008; see also Gibson et al., 2019). In a nutshell, linguistic information originates in a source (speaker, writer), is encoded into symbols (spoken words, written letters) which are transmitted through a channel (sound waves, printed

paper) to a receiver (listener, reader), and is communicated at an average rate (entropy). Crucially, the information contained in each symbol is modulated by predictability (i.e., how surprising the symbol is): when the symbol is predictable, it conveys a low amount of information; when it is less predictable, it conveys a higher amount of information. The core idea for psycholinguistics is that comprehenders will have more difficulty processing words or sentences that are less predictable (i.e., contain high surprisal).

Based on this idea, the Uniform Information Density (UID) hypothesis has emerged as an influential account using information theory to explain how speakers select between alternative variants of a structure (e.g., Jaeger, 2010; Levy & Jaeger, 2006). Key to this hypothesis is the notion of information density – a measure of how much information is communicated per unit of signal when the predictability of this information is taken into account. The claim is that speakers aim for UID in their productions: they prefer to produce a uniform amount of information per unit of linguistic signal in constructing a message and seek to avoid "peaks" and "troughs" in the rate of information provided. The UID hypothesis thus predicts that, in general, speakers will select the longer alternative of a given structure to convey less predictable information and the shorter alternative to convey more predictable information. Figure 1 illustrates how this idea would work in sentences with an optional *that* introducing a new clause. In (a), an object phrase is expected but a new clause is started, causing the first word of the clause not only to carry its typical information content but also to indicate the beginning of the new clause. In this case, the inclusion of the *that* reduces the information contained in the first word, avoiding a peak on the amount of information transmitted per symbol. Conversely, in (b), since a clause is already expected, speakers are predicted to omit the *that*, avoiding the transmission of too little information per symbol. This has been shown for a number of structures including that-deletion in relative clauses (e.g., *this is the friend (that) I told you about*; Ferreira & Dell, 2000) and optional complementizer deletion (e.g., *my boss thinks (that) I'm absolutely crazy*; Jaeger, 2010). The UID hypothesis aims to explain an efficient communication system in which speakers "convey as much information as possible with as little signal as possible," while balancing "the risk of transmitting too much information per time (or per signal), which increases the chance of information loss or miscommunication" (Jaeger, 2010, p. 25).

Virtually, all research on the UID hypothesis to date has focused on speakers' choices in production. However, a similar effect should hold in comprehension: listeners should find it easier to comprehend structures where information density is more uniform. Indeed, this must be the case – there is no point in speakers optimizing the communication channel to provide UID if this is not the preferred way for listeners to comprehend information (Piantadosi et al., 2012). Further, most evidence for the UID hypothesis to date comes from studies of relatively local alternatives, showing the effects of information density in the omission vs. inclusion of optional words (A. F. Frank & Jaeger, 2008; Jaeger, 2010; Levy & Jaeger, 2006). Little evidence is as yet available showing that speakers are sensitive to UID for more complex alternative syntactic encodings.

As far as we are aware, only three studies to date have explicitly investigated the role of UID in the comprehension of more complex syntactic structures. First,

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

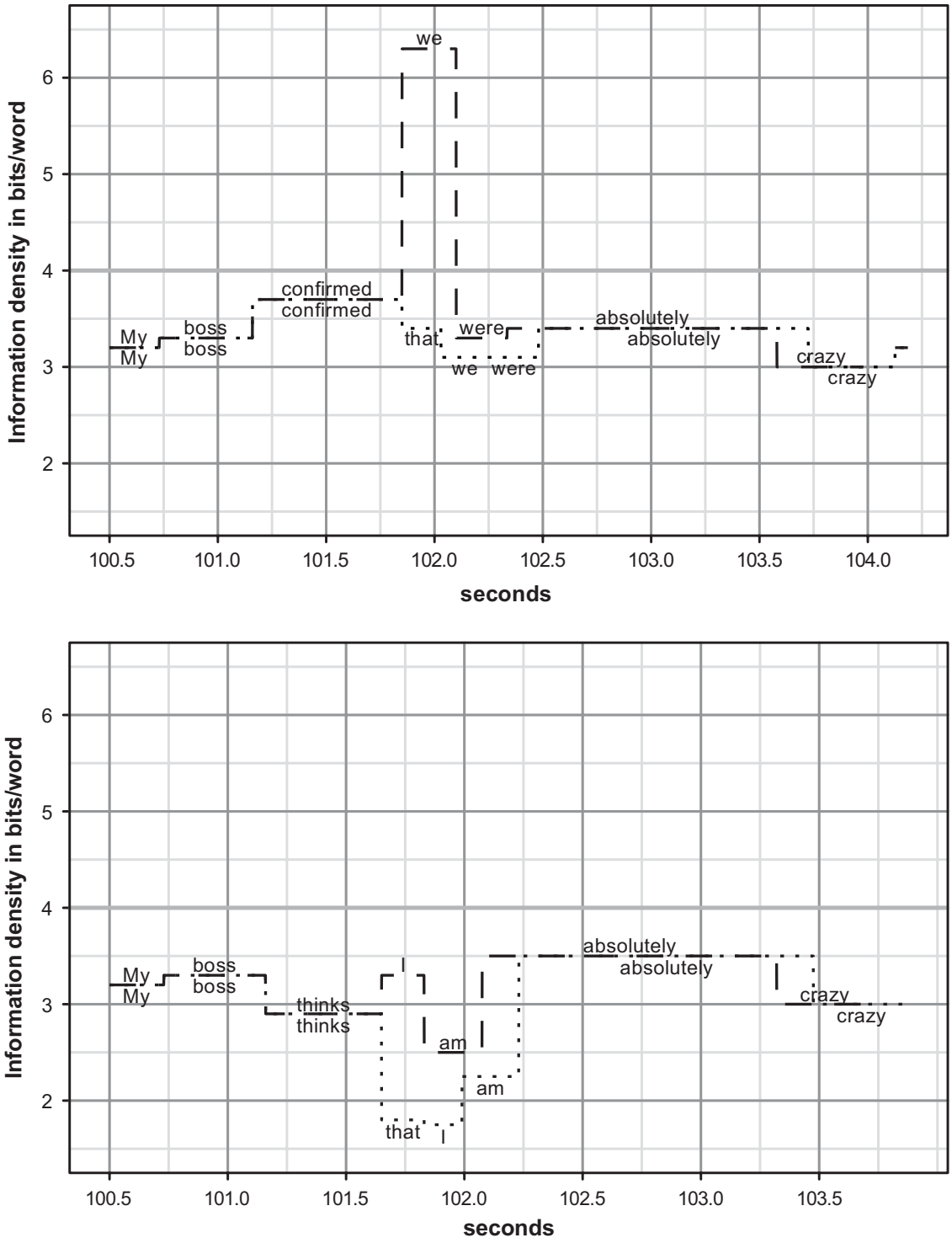


Figure 1. An example of the predictions of the UID hypothesis. Speakers have been shown to prefer using *that* in (a), but not in (b). Reprinted from Jaeger (2010, figure 1), with permission from Elsevier.

Collins (2014) used Mechanical Turk to collect reader preferences on a number of syntactic alternations (e.g., *I looked up the number* vs. *I looked the number up*). They examined to what extent their preferences were in line with the predictions of language parsing models such as Surprisal Theory (Levy, 2008), the UID hypothesis, and Dependency Locality Theory (DLT; Gibson, 2001). They found that none of the theories was able to model the totality of their data and suggested that the UID and DLT are complementary to each other, explaining separate aspects of the parsing

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

system. More recently, Meister et al. (2021) have analyzed a number of different ways in which the UID hypothesis could be operationalized. They analyzed six English corpora containing self-paced and eye-tracking reading times, as well as acceptability judgments, and found that both reading times and reader judgments were best modeled by a super-linear function of each word's surprisal in the sentence. Importantly, this super-linear function meant that the effort is minimal when surprisal is transmitted at a constant rate, supporting the UID hypothesis. Finally, Sikos et al. (2017) assessed participants' preferences for alternative syntactic encodings of noun phrases (e.g., *essay that was carefully written* vs. *carefully written essay*) under different conditions of predictability (e.g., *The journalist published* . . . vs. *The man evaluated* . . .), using the G-maze task in a self-paced reading study in German. Results indeed suggested that participants were sensitive to information density in their comprehension of variants in more complex syntax.

In the present paper, we provide further insight into the effects of syntactic variants on comprehension and its relation to the UID by investigating a different syntactic structure using the more sensitive measure of eye-tracking while reading. We focus on nominal compounds (NC) in English – a structure that denotes one unified concept but consists of a head noun modified by nouns or adjectives, as illustrated in the bolded portion of (1). In the studies reported here, we focus particularly on NCs where the head noun is modified by two or more nouns or adjectives. We contrast these with an alternative structure in which the head noun is modified by prepositional phrases (PP), as in the bolded portion of (2).

1. In some cases, **pharmaceutical market size increase** is driven by the competition in Western countries. (NC)
2. In some cases, **increase in size of the pharmaceutical market** is driven by the competition in Western countries. (PP)

Nominal compounds

NCs[1] are commonly used to convey complex concepts efficiently in a condensed form, serving as "ad hoc names" to refer to new concepts or to further refine or specify existing terms (Bartolic, 1978; Bhatia, 1992; Downing, 1977; Montero, 1996; Pérez Ruiz, 2006; Tobin, 2002; Trimble, 1985; Varantola, 1984). Although relatively rare in conversation, they account for 10–15% of words in academic texts, with the frequency, complexity, and length of the NCs increasing as the level of the text gets more advanced (Biber & Gray, 2010, 2011; Horsella & Pérez, 1991; Salager, 1984; Swales, 1974; Williams, 1984). Their frequency has increased considerably in the scientific register in the last century (see Fig. 2). Complex NCs composed of several words are particularly relevant for the present study because, in the absence of any predictive context, they create a "peak" in information density with respect to the surrounding text in that they convey a large amount of information per unit of communication.

The parsing of NCs is complicated by at least three factors. First, NCs offer fewer signal units (words) than PPs to communicate the same informational content. For example, the bolded information in (1) is conveyed by four words in the NC compared to seven words in the PP equivalent. The NC lacks cues such as

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

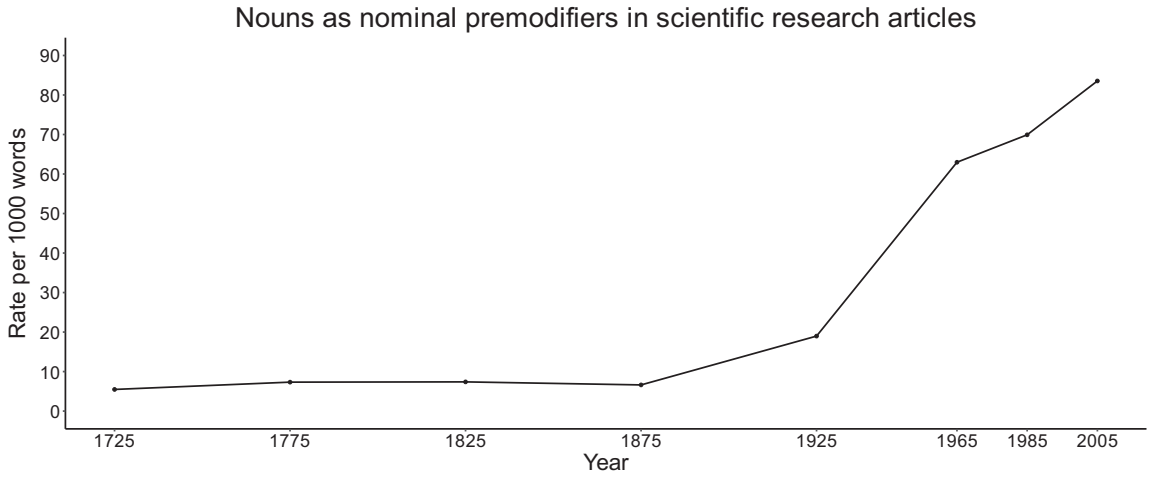


Figure 2. The development of the use of nouns as nominal premodifiers (i.e., forming NCs) through time. Adapted from (Biber & Gray, 2011).

prepositions that elucidate the intended meaning relationships between its components.

Second, NCs in English are head-final so the reader must store the modifiers in working memory during incremental processing and can only integrate them into the parse upon reaching the head noun. Further, since each noun within the NC could potentially serve as the head (e.g., *pharmaceutical market*, *pharmaceutical market size*), the reader could potentially misanalyze the phrase at each word and cannot fully parse the NC until a verb signals its completion. In the PP variant, however, the head noun begins the noun phrase so the reader can immediately integrate the modifiers without holding them in working memory. Several theories of parsing make clear that locality – the distance between two elements that are dependent on each other – plays an important role in integration difficulty at the head (e.g., Gibson, 2001; Vasishth, 2010).

Third, the string of words in the NC usually does not enable the reader to quickly grasp which of the large number of possible underlying syntactic and semantic relationships a particular NC conveys. Noun compounds can convey the same semantic information as a number of alternative syntactic structures and relationships (e.g., verb-argument relationships, relative clauses, prepositional phrases), but these are not evident in the NC itself (Lees, 1960, 1970; Levi, 1973, 1978; Limaye & Pompian, 1991; Pérez Ruiz, 2006; Warren, 1978). Even the basic branching structure within longer NCs is typically not discernible without context; three-word NCs can be left-branching (e.g., *health service employee*), right-branching (e.g., *head copy boy*), or ambiguous between the two (e.g., *steel bridge foundation*) (Kvam, 1990). A wide variety of semantic relationships can also hold between words in an NC; lists of the possible relationships for two-word NCs proposed in the literature extend from four relationships (Granville Hatcher, 1960) to over 100 (Brekle, 1976). Further, many NCs are ambiguous between two or more potential semantic interpretations. For example, *translator writing system* could be understood as either Purpose (writing system for the translator) or Source (writing system of translators) (Montero, 1996, p. 66). Because the syntactic and semantic relationships are not overtly expressed within the NC itself, they must be actively reconstructed online by the reader based on pragmatic

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

and contextual information, which leads to potential processing difficulty and delay (Bauer & Tarasova, 2013; Berg, 2016). In contrast, the additional words in PPs such as prepositions and verbs facilitate awareness of syntactic and semantic relationships during online processing.

In summary, the aforementioned literature has considered NCs from a number of different but converging perspectives strongly suggesting that NCs should be harder to process than PPs. In this paper, we assume that all these perspectives are subsumed under the notion of information density.

But are NCs really harder to process? Given the frequent use of complex NCs in academic texts combined with the potential comprehension difficulties they present, it is surprising that relatively little is known about how complex NCs are processed by readers. Most of the research on the processing of NCs has focused on two-constituent compounds and has examined compounds in isolation using tasks such as lexical decision, sense-nonsense judgment, and masked priming (e.g., Estes & Jones, 2006; Gagné & Shoben, 1997; Gagné & Spalding, 2009; Schmidtke et al., 2016). Only a few studies have looked at longer compounds (e.g., de Almeida & Libben, 2005; Inhoff et al., 2000; Krott et al., 2004) or at compounds in actual sentences (e.g., Kuperman et al., 2008), but not from the point of view of information density. Several off-line behavioral studies with longer NCs have shown that there is a considerable error and lack of consistency across individuals in paraphrasing NCs (e.g., Geer et al., 1972; Gleitman & Gleitman, 1970; Olshtain, 1981; Williams, 1984), selecting appropriate definitions for NCs (e.g., Gleitman & Gleitman, 1970; Limaye & Pompian, 1991), and translating NCs to other languages (e.g., Carrió Pastor, 2008; Carrió Pastor & Candel Mora, 2013). Overall, this literature strongly suggests that NCs are challenging to process, but provides no information about the comprehension of these constructions in real time. Further, none of these studies has focused directly on information density, comparing the processing of NCs to other less dense structures.

The present study[2]

Therefore, in the studies reported here, we examine the processing of NCs vs. PPs in real time using eye-tracking while reading, in the absence of any preceding context that would allow the meaning of the critical segment to be predicted. We report two separate experiments in English where L1 participants read sentences containing NCs and PPs of different lengths. In the first experiment, sentences contained critical structures (NC or PP) of lengths 4 and 6. In the second, which controls for a number of limitations identified from Experiment 1, shorter critical structures of lengths 3 and 4 were used.

We predict that participants will experience more processing difficulty with the informationally dense structure (NC) than with the less dense structure containing roughly the same total information content (PP). Since the target structures vary considerably in length, we do not measure the time participants spend looking directly at them, but instead operationalize this difficulty in two indirect ways. First, we count the number of times participants regress towards the critical structure (i.e., toward the NC or the PP), predicting a higher number of regressions in sentences containing NCs (vs. PPs) and longer structures (6 vs. 4 words in Experiment 1, 4 vs. 3 words in Experiment 2). Second, we measure the time participants spend looking

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

at the text segments *following* the critical structure, in a fashion similar to other studies in the eye-tracking literature (e.g., Christianson et al., 2017; Jared & O'Donnell, 2017; Paape & Vasishth, 2022; Pickering & Traxler, 1998). Since these (subsequent) text segments do not change between conditions, we can say that, if participants gaze for a significantly different amount of time in one condition, then this can only be attributed to the experimental manipulation of the critical structures preceding them. In order to evaluate the time course of this processing difficulty, we extract eye-tracking measures from the data reflecting both early and late effects, and predict that participants will spend more time looking at the segments following the critical segment after reading an NC than after reading a PP and after reading a longer structure than after reading a shorter structure. Since no other study has used eye-tracking to investigate online NC processing, and in an attempt to be statistically conservative, we a priori selected one early and one late reading measure.

Experiment 1

Method

Participants

Participants were 31 L1 English speakers recruited from the communities of the University of Alberta (Canada) and the Technische Universität Kaiserslautern (Germany). One participant was excluded from the analysis due to exceptionally noisy eye-tracking data (see "Eye-movement analysis" below). We report below the data of the remaining 30 participants (mean age: 25.2, SD: 8.31). No participants were exposed to a second language before the age of 5 years and did not regularly use any other language in their daily life. All had normal or corrected-to-normal vision. Participants were compensated with payment or course credit.

Materials

Eye-tracking reading task

Participants read sentences containing either a nominal compound (NC) or the same noun phrase in the form of a head noun followed by prepositional phrases (PP). The number of content words in the NC/PP was manipulated: there were 12 critical items containing four (4) content words, and 12 critical items containing six (6) content words. The relationship between the ordering of the content words in the NC and in its corresponding PP was always the same. For 4-length items, the NC order was N1-N2-N3-N4, while the PP order was N4-P-N3-P-N1-N2. For 6-length items, the NC order was N1-N2-N3-N4-N5-N6 and the PP order was N6-P-N5-P-N3-N4-P-N1-N2. The 4-length and 6-length items were completely different, that is, there is no overlap between the words used in 4-length and in 6-length items. The variables Phrase Type (NC vs. PP) and Length (4 vs. 6) were combined to yield four conditions (see Table 1 for example items; see full set of critical items in Appendix A). In order to reduce cognitive load, and in order to ensure that the materials were equally familiar

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

Table 1. Example sentences for each experimental condition in Experiment 1

Condition	Preamble	Critical region	Region of interest 1	Region of interest 2	Wrap up
4-NC	In present times,	health insurance economy effects	are researched	by the analysts	of financial institutions
4-PP	In present times,	effects of the economy on health insurance	are researched	by the analysts	of financial institutions
6-NC	In current times,	United States factory employee insurance costs	are decreased	by the changes	in union policies
6-PP	In current times,	the cost of insurance for factory employees in the United States	are decreased	by the changes	in union policies

Condition	Comprehension Questions
4-NC/PP	Financial institution analysts research health insurance
	X: yearly costs M: economy effects
6-NC/PP	Union policy changes decrease employee insurance costs in
	X: the United States M: France

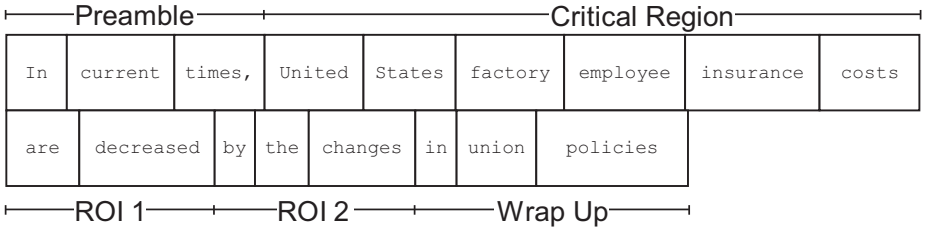


Figure 3. An example trial reconstructed from the data collected in Experiment 1. (The reconstruction's font size does not reflect the exact font size of stimulus presentation – note the varying distance between words.) The boxes around each word indicate the interest area associated with that word: fixations inside a given interest areas are treated as fixations on its word. The boxes and the indication of the sentence regions were not visible to readers during the experiment.

for all participants, the sentence content was kept theme constant, always focused on economics and business (none of the participants were studying economics).

The critical sentences were composed of five regions (see Table 1): a preamble containing three words and ending with a comma; a critical region containing either an NC with 4 or 6 words or an equivalent PP; a passive construction (ROI1) composed of the auxiliary "is" or "are" and a participle; the agent of the passive construction (ROI2) containing a single noun and introduced by "by the"; a final "wrap up" region. The critical segment was positioned early in the sentence to avoid influence of any preceding context on comprehension. The number of characters and syllables in the content words of the NC and PP was controlled (see Supplementary Materials). The introductory phrase and critical segment were on the first line and the other three segments on the second line (see Fig. 3 for a reconstruction of how the items were presented on the screen).

Oxford Placement Test Part 1 (OPT)

Participants' English proficiency was assessed using the OPT. It consisted of a series of sentences containing 50 gaps at which three possible completions were presented and only one was grammatically correct. The participant's task was to choose the correct option.

Language Background Questionnaire (LBQ)

In order to ensure that participants were native speakers, they also filled out a language background questionnaire responding to questions about the languages they speak, the situations in which they use them, how well they are able to use them, and the people with whom they use them.

Digit span test (DST)

We assessed participants' working memory using a digit span test (WM; see Wambach et al., 2011). Participants saw sequences progressing from 2 to 9 digits and had to recall the correct order for at least two sequences of a given length to advance to the next length. A participant's final digit span was the longest length for which they correctly answered at least two sequences. (see Supplementary Materials).

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

Design

The task consisted of 68 trials: 4 practice trials, 24 critical items (12 of each length), and 40 fillers. The fillers had similar syntactic complexity and length to the critical sentences, but did not contain any NCs longer than two words. The design is partially factorial: phrase type (NC vs. PP) was manipulated within participant and within item, and length (4-length vs. 6-length) was manipulated within participant and between items. We additionally consider the interaction between type and length (see "Eye-movement analysis" below for details). The NC and PP versions of the twelve 4-length and the twelve 6-length critical items were separately assigned to two counterbalanced lists, such that each list contained six NC and six PP variants of each length. The presentation order was randomized per participant. After each item, a comprehension question was displayed, and the participant pressed the letters X or M to answer. Questions after critical items probed the head noun (see Table 1).

Apparatus

Stimulus presentation was programmed with the Experiment Builder software from SR Research, and eye movements were recorded with an Eyelink 1000, sampling at 1000 Hz. The eye tracker recorded movements from the right eye, though presentation was binocular. Participants viewed the stimuli on a color monitor at a resolution of 1280 × 1024 and a distance of approximately 100 cm, using a chin rest to stabilize their head. The sentences were presented using the font Courier New. Eyes were calibrated and validated at the beginning of the study, halfway through, and as needed throughout the study.

Procedure

After giving informed consent, participants completed four tasks in the following order: the Oxford Placement Test Part A (OPT; Allan, 2006) assessing English proficiency, a language background questionnaire, the DST, and the eye-tracking task. For the eye-tracking task, participants were informed that they would read English sentences from economics texts while having their eye movements recorded, and after each sentence, they would have to answer a question about the sentence they just read. They were also presented with written instructions on the screen and given the chance to ask questions. The eye tracker was then calibrated, and the participant read the four practice trials. If needed, a new calibration was then performed. Finally, the two blocks of 34 sentences each were presented. Each item started with a drift correct that corresponded to the sentence's first letter. After reading the sentence, participants pressed the spacebar to advance to the comprehension question. Participants were told to read normally with no time limit.

Eye-movement analysis

Prior to the analysis of the eye-movement data, trials with incorrectly answered comprehension question were discarded (see Table 2). As is common with eye-tracking data (Hornof & Halverson, 2002; Zhang & Hornof, 2011; Blignaut et al., 2014; Zhang & Hornof, 2014), fixations in many trials contained systematic errors

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

Table 2. Participant mean accuracy per condition of Experiment 1. Incorrect trials were discarded from the eye-movement analysis

Language	Type	Length	Mean accuracy (Proportion)	SD	# Trials correct	Total trials
English	NP	4	0.922	0.114	165	179*
English	NP	6	0.961	0.095	173	180
English	PP	4	0.967	0.068	174	180
English	PP	6	0.950	0.089	171	180
			Trials kept		683	94.99%
			Trials discarded		36	5.01%

*: Recall that 1 trial was removed before analysis because it had fewer than 5 fixations.

that could be easily corrected. In order to account for this error while avoiding human bias, we applied an automatic correction procedure on all fixations of all trials prior to analysis (see Supplementary Materials). We then visually inspected each trial looking for trials containing fewer than five fixations (1 trial) and any trial that was clearly noisy either before correction (i.e., the algorithm could not possibly have meaningfully corrected the data) or after correction. All noisy trials were concentrated on a single participant, who, as mentioned in the Participants section, was thus removed entirely from the data. All statistics reported below are based on the remaining corrected data.

For the analysis, we a priori chose one early reading measure – **First Pass Duration (FPD)** – and one late reading measure – **Total Duration (TD)** – for our analyses. First Pass Duration was defined as the summed length of all fixations on a given region for the first time the participant arrived at that region, before moving past it.[3] Total Duration was defined as the summed length of all fixations on a given region. These measures were extracted from the corrected fixations for each of ROI1 and ROI2 using the Get Reading Measures script (Dan, 2020). In addition, we extracted a count of all the **regressions onto the critical region (Reg2CR)**. A fixation counted as a Reg2CR if it satisfied two conditions: (1) it was located inside the critical region, and (2) the previous fixation was located inside a subsequent interest area. Note that this definition will include regressions performed completely *inside the critical region*. For example, given the critical region *pharmaceutical market size increase*, a fixation on *pharmaceutical* will count as a regression if its previous fixation was performed on any of *market*, *size,* and *increase*. We chose to include these cases because the critical region of our items was quite long (cf. *increase in the size of the pharmaceutical market*).

We trimmed the data in order to avoid extreme reading measure values. For each ROI, we trimmed extreme values by discarding trials whose FPD was below 80 ms or greater than 1000 ms, and trials whose TD was below 80 ms or greater than 2000 ms (see Table 3). Because we analyzed five reading measures (two duration measures over two ROIs, plus the regression count), we applied a Bonferroni correction resulting in an alpha threshold of $0.05/5 = 0.01$ (see von der Malsburg & Angele, 2017).

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

Table 3. Number (percentage) of trials trimmed before the analysis of each reading measure of Experiment 1

	Region of interest 1			Region of interest 2		
	Discarded		Kept	Discarded		Kept
FPD	112	(16.40%)	571	60	(8.78%)	623
TD	32	(4.69%)	651	54	(7.91%)	629

The extracted duration measures were submitted to generalized mixed models (GMM) with a Gamma distribution and an *identity* link function to account for the skewed nature of duration measures (see Lo & Andrews, 2015), using the lme4 package (Bates et al., 2015) in R (R Core Team, 2013). Results include *p*-value estimates from the lmerTest package (Kuznetsova, Brockhoff, Christensen, et al., 2017). Given that we had clear hypotheses, fixed effects included Phrase Type (NC/PP), and Length (4/6), both of which were sum-coded, as well as two predictors (OPT score and DST score), which were scaled and centered to reduce collinearity. Also included were random effects for subjects and items, which were maximally specified (Barr et al., 2013)[4]. Full model estimates and values are provided in Appendix C.

The number of regressions onto the critical region is a type of count data and thus is not expected to follow a normal distribution. Therefore, we fitted a GMM, using a Poisson distribution with a *log* link function. The model was maximally specified, in the same way as the duration measures.

Results

Comprehension accuracy was high as can be seen in Table 2 and will not be considered further. Participants scored a mean of 93.45% (SD: 4.96) on the OPT and a mean of 6.64 out of 9 (SD: 1.25) on the DST. Table 3 shows the number of discarded trials due to trimming of extreme reading measure values, and the number of trials used in the final eye-movement analyses.

Figures 4 and 5 show the reading measures for regions of interest 1 and 2, respectively. The GMMs revealed no effect of Length or Type in any of the duration models analyzed (see Tables C1, C2 for First Pass Duration in ROI1 and ROI2, respectively, and Tables C3 and C4 for Total Duration in ROI1 and ROI2, respectively). However, in the analysis of Total Duration (see Fig. 7a and 7b), we did find an effect of the DST score on both ROI 1 ($t = -4.367$, $p < .001$) and ROI2 ($t = -4.816$, $p < .001$), with higher WM scores leading to shorter durations.

Figure 6 shows the distribution of the number of Reg2CR by condition. We do not present mean and standard deviation because the distribution is notably not normal. Trials with length 6 had a significantly larger amount of Reg2CRs ($z = -4.645$, $p < .001$). Trials containing an NC had significantly fewer Reg2CRs than trials containing PP ($z = -5.312$, $p < .001$). In addition, English proficiency was a significant predictor of the number of Reg2CR ($z = 2.681$, $p < .007$; see Fig. 7c), but not in the expected direction: higher proficiency led to more Reg2CRs

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

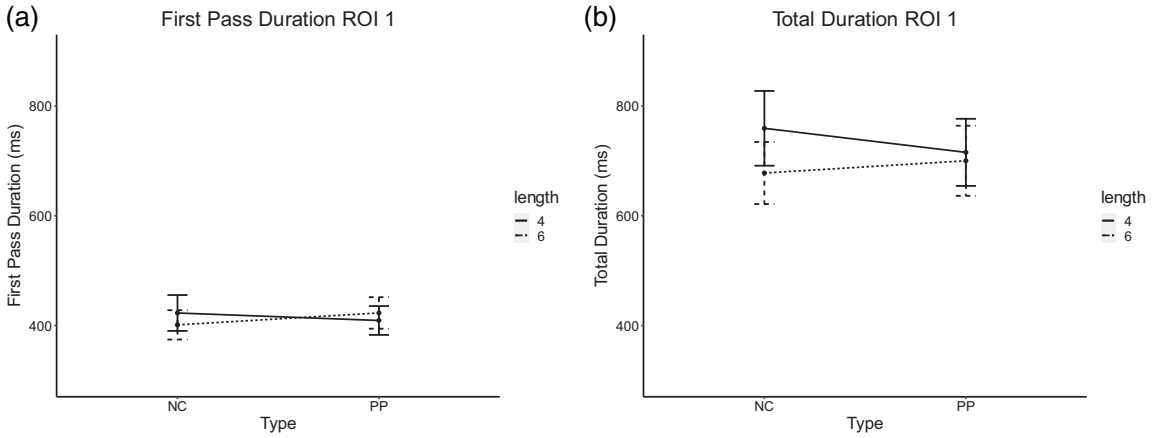


Figure 4. Experiment 1: Reading measures in Region of Interest 1. Error bars represent 95% confidence intervals.

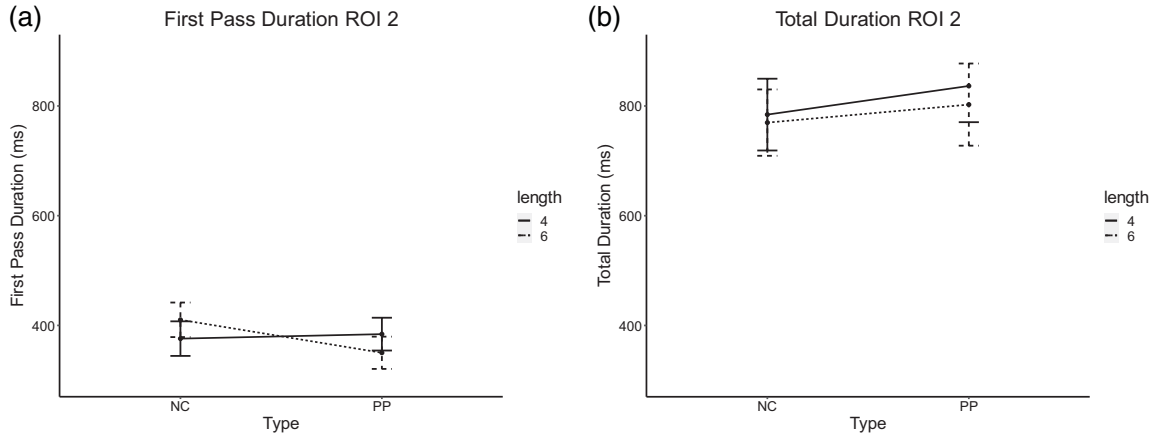


Figure 5. Experiment 1: Reading measures in Region of Interest 2. Error bars represent 95% confidence intervals.

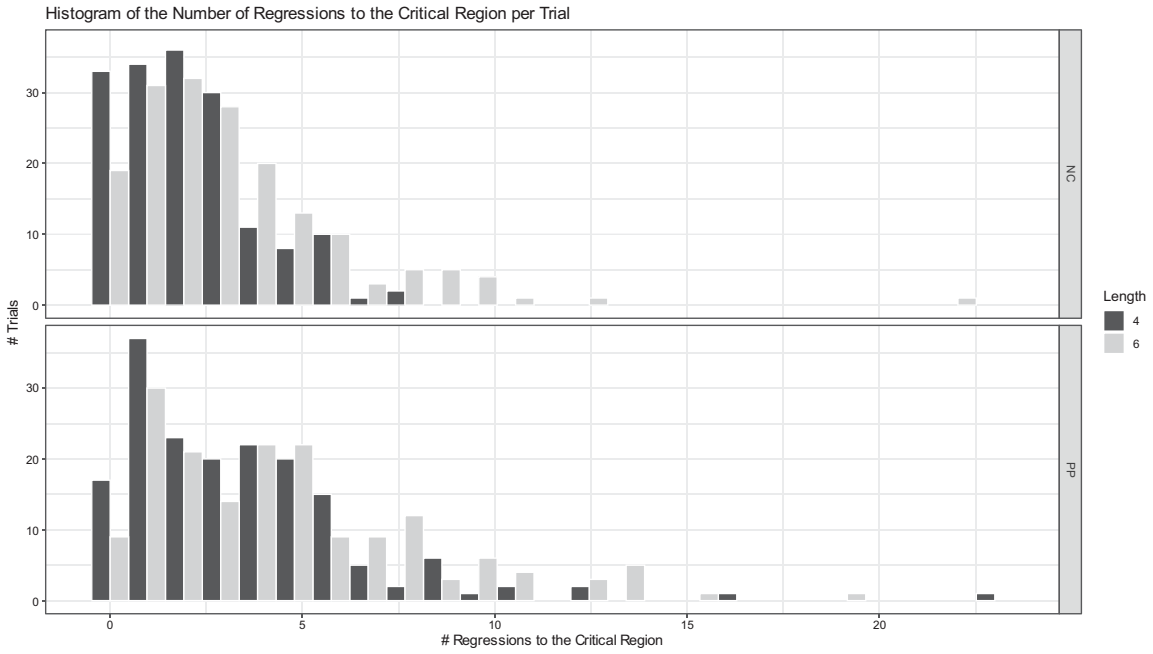


Figure 6. Experiment 1: Regressions onto the critical region.

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

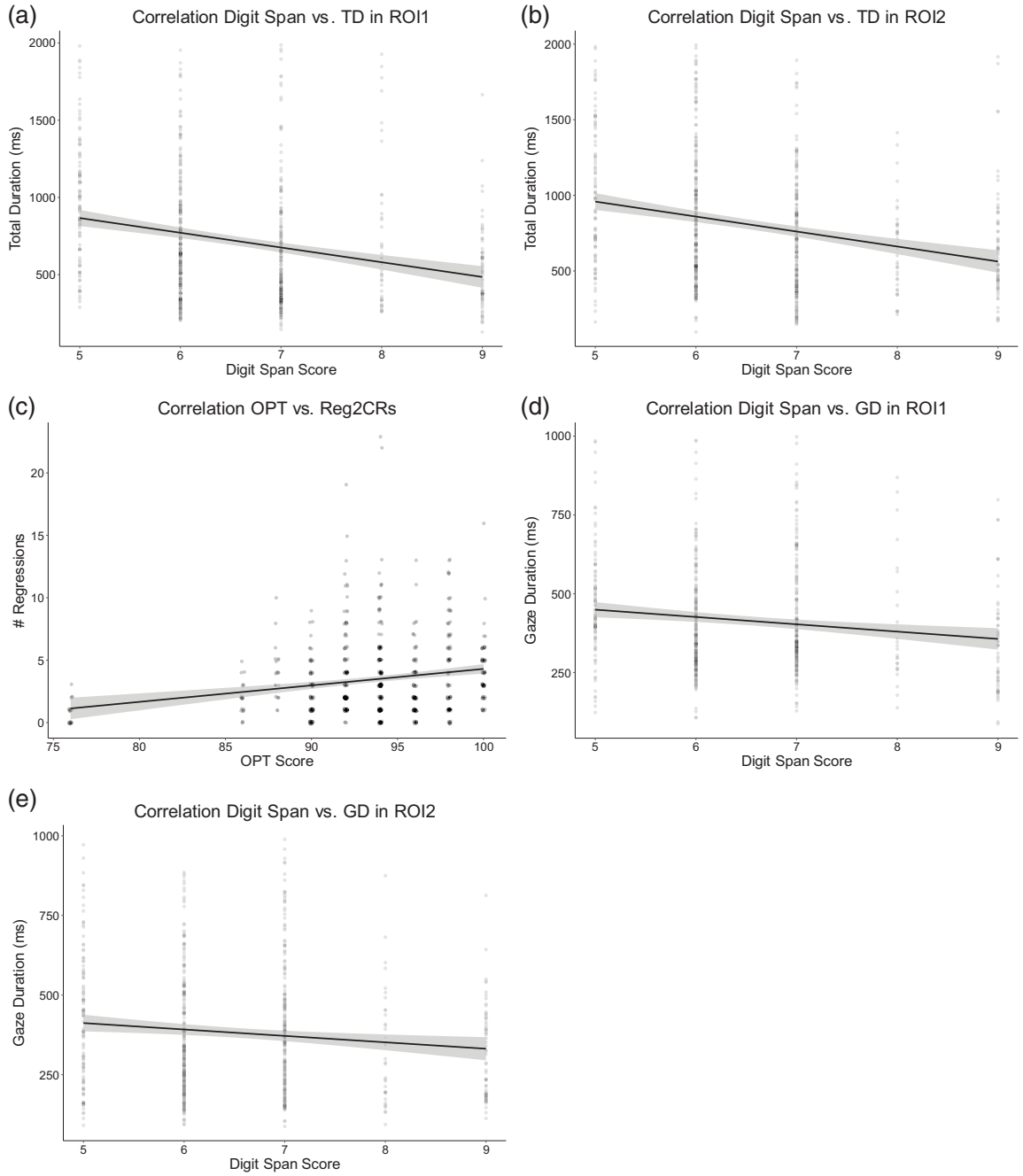


Figure 7. Experiment 1: The relation between the reading measures and some of the covariate data in our study. We use transparent data points in order to indicate their density. The covariates in (a), (b), and (c) were found to be significant predictors of the reading measures they are presented with. The covariates (d) and (e) are referred to in the Discussion. Gray shading indicates 95% confidence intervals for the regression line.

than lower proficiency (see Table C5). This effect was driven by a single outlying participant with very low proficiency score (see Fig. 7c) and became non-significant when the participant was removed.

Discussion

In this experiment, we investigated the reading difficulty evinced by NCs compared with that caused by PPs. In particular, we examined structures composed of either

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

four or six content words. Following the UID hypothesis, we predicted that the longer structures would cause more difficulty than the shorter ones and that NCs would be harder to process than PPs. In addition, we predicted that this difficulty would be modulated by participant individual differences such as WM (measured by the DST) and English proficiency (measured by the OPT). To measure this difficulty, we extracted the number of regressions onto the target structures, as well as two duration measures in the two regions following the target structure, reflecting early and late processing effects. Our assumption was that more regressions and longer duration measures would indicate more processing difficulty.

Surprisingly, we found no effect of either length or type for any of the duration measures, neither in the segment immediately following the target structure nor in the subsequent region. It is not clear why this is the case. If NCs *are* more difficult to process, constituting peaks of information density, then this difference in difficulty was not revealed by either early or late reading measures. We surmise that this may have occurred for at least three reasons: because of the a priori choice of reading measures, because this difficulty did not spill over to the subsequent sentence segments as we initially expected, or because of the large variance observed in these measures.

Turning to the regression data, we did find an effect of length indicating, as predicted, that 6-length structures led to more regressions than 4-length structures. These results support the UID hypothesis under the assumption that 6-length structures do generally involve higher peaks of information than 4-length structures. We also found an effect of type on the number of Reg2CR, but this effect indicated a higher number of regressions in the PP condition, contrary to our predictions. We return to this matter in the General Discussion.

Finally, turning to our covariates, proficiency did not affect reading measures, a result we also return to in the General Discussion. Conversely, DST did significantly predict TDs (in ROI1 and ROI2), as expected, but surprisingly not FPDs. As Fig. 7d and 7e suggest, however, it is possible that the effect of DST was too small for the power of the current experiment.

In sum, speakers did not fixate longer on the text segments following the critical regions, regardless of the length or the type of structure present in the critical region, but their fixations were generally modulated by their WM abilities, at least when considering the total time spent looking at these regions. Finally, participants regressed more after longer structures, and (contrary to expectations) after PPs.

Note, however, that the experiment reported above has several limitations that may have affected the results. First, the definition of NCs presented above allows for adjectival modifiers to be interleaved with nouns (e.g., ***modern*** *era* ***general*** *election campaign corruption*). Since in these contexts it is clear that the adjectives are modifying a subsequent noun, they might have helped the participants in predicting that the next word was also a continuation of the NC, producing some facilitation. From an information theoretic perspective, this would amount to modifying the probability distribution of the next word such that all nouns are a little more likely (are less surprising or informational) and all other words are a little less likely (more informational), ultimately smoothing the information density peak caused by the NC.

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

Second, assuming that, in the absence of external context, items with more content words are more informationally dense, we would expect longer items to evince more difficulty than shorter ones. However, the design used in Experiment 1 was not fully within-items, so that any effect of length we have found in Experiment 1 may be an artifact of a difference in the items themselves, rather than a real effect. In other words, it would be ideal if the items were designed in such a way that longer items contained the same words as the shorter items, so that any difference between shorter and longer items is unequivocally attributable to the differing words in these items.

Third, the content words used in many of the items in Experiment 1 can be used as both a noun and a verb. This may have caused garden path effects, introducing a confound in the experiment results. For example, when reading the item *United States factory employee insurance costs*, participants could have processed the last word (*costs*) as either a verb (therefore expecting an object afterward) or a plural noun (therefore expecting a verb afterward). Specifically in this case, the NC is already quite long, and this could have made the participants even more likely to expect it to be a verb, and not a noun.[5]

Fourth, because the NCs used in Experiment 1 were inspired by real economics and business texts, they included certain collocations[6] (e.g., United States). These collocations might have helped participants in deciphering the structure of the NC, reducing the difficulty perceived when processing them as compared to PPs. That is, if the items were controlled for these collocations, we would expect a starker difference between the difficulty perceived in processing NCs and that of processing PPs. In information theoretic terms, collocations are not informationally dense: after the first word is encountered, the subsequent words are very much expected and contain little information. Hence, NCs composed of collocations may be even *easier* than PPs if they are common enough (e.g., *heart rate variability vs variability in rate of the heart*).

Finally, Experiment 1 included a DST with the purpose of assessing WM ability.

The rationale was that participants with better WM abilities would have more resources available and therefore would be less affected by the peak of information caused by NCs. Hence, WM would act as a predictor of the number of regressions and of the time spent in the ROIs 1 and 2 after an NC. However, WM is a complex construct, composed of (among others) a phonological loop involved in the processing of language stimuli, and a visuospatial sketchpad involved in the processing of visual stimuli (see, e.g., Baddeley, 2011 for a review), and it is not clear which of these subcomponents the DST taps into.

In Experiment 2, we use the same paradigm, but address each of the limitations just described.

Experiment 2

In Experiment 2, each NC is composed solely of nouns that normally cannot be used as verbs, and do not contain collocations. In addition, to allow for a fully within-items experiment design, we use NCs of length 3 and 4, which were constructed such that those composed of 4 content words are an extension of the 3-length NCs.

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

As discussed earlier, the UID hypothesis predicts NCs to lead to more difficulty than PPs, since they are used at the beginning of the sentences, without any helpful context, producing a peak of information. For similar reasons, we also expect longer constructions to lead to more difficulty than shorter ones: without any context or collocations to guide the participants' expectations, longer structures should correspond to starker peaks of information.

Instead of a DST, in Experiment 2 we use two WM tasks: a verbal and a non-verbal task. Both tasks are serial order reconstruction tasks (SORT; Jones et al., 1995), where participants are shown a sequence of items one-by-one and are later asked to select them in the order in which they were displayed. Given the direct relationship between the phonological loop and language processing, we expect that the two tasks will not be related to each other, and that the verbal SORT will predict reading behavior, but that the visual SORT will not.

We measured participants' English proficiency with the OPT and assessed the quality of lexical representation (cf. Perfetti & Hart, 2002) with a misspelling identification task (MSIT). We use the two tasks because spelling has been shown to predict variance independently of vocabulary and reading comprehension measures in priming and eye-movement data (see e.g., Andrews et al., 2020).

Method

Participants

The participants were 39 English native speakers, all of whom were students at the University of Alberta. They had normal or corrected-to-normal vision and were compensated with course credits. Out of the 39, 13 participants reported substantial exposure to a language other than English before the age of 5 and were therefore discarded from further analysis. The data of the remaining 26 participants, aged between 17 and 28 (mean age: 20.2, SD: 3.29), are reported below.

Materials

Eye-tracking reading task

As in Experiment 1, participants performed a reading task in which they read sentences containing either an NC or a PP. Each sentence was divided into the same five segments as Experiment 1: a semantically neutral introductory phrase, the critical segment (NC or PP), the first and second regions of interest (ROI1 and ROI2, respectively), and the final segment (not analyzed).

The 4-length NCs were composed exclusively of nouns and were constructed so that they were identical to the 3-length NCs except for the first word (e.g., inflation constraint action vs. *currency* inflation constraint action). The items were reviewed by three native speakers who were aware of the purposes of the experiment and who additionally checked for collocations. None of the nouns repeats across items.

We additionally controlled the critical items in a number of ways. The preamble was again composed of 3 words, but this time started exclusively with the preposition "in." The critical segment was always preceded by "the." The passive construction (ROI1) used exclusively the auxiliary "is." The nouns composing the NCs were controlled for length, such that all 3-length NCs had between 24 and 29

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

characters and all 4-length NC had between 31 and 36 characters. The additional word inserted to create a 4-length NC was always between 6 and 10 characters long. See Table 4 for example items, and Appendix B for a list of all items.

Oxford Placement Test Part 1 (OPT)

This was the same as in Experiment 1.

Language Background Questionnaire (LBQ)

This was the same as in Experiment 1.

Misspelling Identification Task (MSIT)

The MSIT was performed in addition to the OPT. In this task, participants received a list of 215 words, 50 of which were incorrectly spelled. Their task was to circle all incorrectly spelled words present in the two pages. The scoring was based on the LexTale (Lemhöfer & Broersma, 2012) scoring formula. See the Supplementary Materials for scoring details and all task words.

Working memory tasks

Participants also performed a visual and a verbal serial order reconstruction task (SORT; Jones et al., 1995). Both tasks measured participants' WM abilities, but tapped into a different memory component. In each trial, participants saw a sequence of letters or dots, depending on the task. After the sequence, all its elements were shown again and the participant was tasked with clicking on the elements in the order in which they had appeared (3 practice and 20 critical trials). See Supplementary Materials for details.

Design

The eye-tracking task was composed of a total of 4 practice sentences, 40 filler sentences, and 28 critical sentences. They were divided into three blocks: a set of 4 practice sentences, presented in a random order, followed by a randomized set of 68 trials divided into two blocks of 34 trials containing both filler and critical sentences, and separated by a short break. The experimental items were randomly assigned to four lists forming a 2 × 2 Latin square design (type: NC vs. PP; length: 3 vs. 4) such that each participant only saw one version of each sentence.

After each item, participants answered a comprehension question by pressing the letters X or M. In order to avoid disrupting participants' typical reading behavior, the questions were designed to be easy.

Apparatus

Stimulus presentation was programmed with the Experiment Builder software from SR Research, and eye movements were recorded with an Eyelink 1000 Plus, sampling at 500 Hz. The eye tracker recorded movements from the right eye, though presentation was binocular. Participants viewed the stimuli on a 20-inch color

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

Table 4. An example critical item of Experiment 2

Condition	Preamble	Critical region	Region of interest 1	Region of interest 2	Wrap up
3-NC	In present times,	the inflation constraint action	is implemented	by the Board	of the National bank
4-NC	In present times,	the currency inflation constraint action	is implemented	by the Board	of the National Bank
3-PP	In present times,	the action for the constraint of inflation	is implemented	by the Board	of the National Bank
4-PP	In present times,	the action for the constraint of inflation of the currency	is implemented	by the Board	of the National Bank
Comprehension question		The National Bank board implements			
		X: inflation constraint actions			
		M: deflation constraint actions			

monitor at a resolution of 1280 × 1024 and a distance of approximately 90 cm, using a chin rest to stabilize their head. The sentences were presented using the font Courier New. Eyes were calibrated and validated at the beginning of the study, halfway through, and as needed throughout the study.

Procedure

After signing the informed consent form, participants were given the language background questionnaire. The OPT was then administered, followed by the MSIT.

The participant then was asked to sit in front of the eye tracker and instructed about the eye-tracking reading task, whose procedure was virtually the same as that of Experiment 1. As in Experiment 1, sentences were presented so that the preamble and the critical region appeared on the first text line, and the rest appeared on the second line.

After finishing the eye-tracking task, the participant moved on to another computer where the two short-term memory tasks were performed. The visual serial order reconstruction task was performed first, followed by the verbal SORT. This ordering was chosen to prevent participants from being influenced by the verbal task when performing the visual one: if the verbal task were performed first, participants could be led to use a verbal strategy to perform the visual task, yielding spurious results. Before each task, the task instructions were both explained by the experimenter and presented on the screen for the participant to read.

Eye-movement analysis

The analysis is roughly the same as in Experiment 1. Trials with incorrect comprehension question responses were discarded (see Table 5), and the fixation correction resulted in the removal of 7 additional trials (see the Supplementary Materials for examples of these cases).

From the resulting data, we extracted FPD and TD for each of ROI1 and ROI2, as well as the number of Reg2CR. Trimming and Bonferroni correction were applied as in Experiment 1.

The extracted duration measures were submitted to separate GMMs with a Gamma distribution and an *identity* link function. Fixed effects included two predictor variables for proficiency (OPT score and MSIT score) and two WM scores (verbal SORT and visual SORT), all of which were scaled and centered. We also added Length (3/4) and Phrase Type (NC/PP) as fixed effects, both of which were sum-coded. The random effects structure of the models was maximally specified.[7]

Regression counts were analyzed in an analogous manner, with a maximally specified GMM using the Poisson distribution and the *log* link function.

Results

Comprehension accuracy was generally high, as reported in Table 5, and, as in Experiment 1, was not considered further. Participants also scored high both in the OPT (46.440 ± 2.583 out of 50) and in the MSIT (82.94 ± 7.636 out of 100). The amount of trimming for each duration measure is reported in Table 6.

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

Table 5. Participant mean accuracy per condition of Experiment 2 after data removal. Incorrect trials were discarded from the eye-movement analysis

Language	Type	Length	Mean accuracy (Proportion)	SD	# Trials correct	Total trials
English	NP	3	0.928	0.102	168	181
English	NP	4	0.938	0.097	170	181
English	PP	3	0.973	0.057	177	182
English	PP	4	0.908	0.131	161	177
			Trials kept		676	93.76%
			Trials discarded		45	6.24%

Table 6. Number of trials trimmed before the analysis of each reading measure of Experiment 2. Numbers inside parenthesis indicate the percentage of the total trials

	Region of interest 1		Region of interest 2	
	Discarded	Kept	Discarded	Kept
FPD	98 (14.50%)	578	49 (7.25%)	627
TD	19 (2.81%)	657	38 (5.62%)	638

Figures 8 and 9 show the results for ROIs 1 and 2, respectively. No effect was found for TD in either ROI1 or ROI2. For FPD in ROI2, we found an effect of Type, with trials containing an NC evoking significantly longer FPDs than trials containing a PP ($t = 2.974$, $p = .003$). In addition, visual WM score was a significant predictor of FPD in ROI1 ($t = 3.540$, $p < .001$) and in ROI2 ($t = 2.684$, $p = .007$), with longer FPDs associated with better visual WM abilities (see Fig. 10). This effect was driven by a single participant with a very high visual WM score. Removing the participant made the effect no longer significant (see Fig. 10c and 10d). See Tables D1 and D2 in Appendix D for FPD in ROIs 1 and 2, respectively, and Tables D3 and D4 in Appendix D for TD in ROIs 1 and 2, respectively.

Figure 11 shows a histogram of the number of Reg2CR. Again, we do not show mean and standard deviations because the distribution was clearly not Gaussian. We found a significant effect of Length ($z = -4.276$, $p < .001$), indicating a higher Reg2CR count for trials containing structures with length 4 than those with length 3. In addition, NCs evoked significantly fewer Reg2CR than PPs ($t = -9.821$, $p < .001$). See Table D5.

As expected, visual WM scores (mean: 4.526, SD: .969) and verbal WM scores (mean: 4.631, SD: 1.380) were not significantly correlated (correlation: 0.129, $t = .638$, $p = .530$).

In order to verify that there was no large correlation between the model variables that could lead to a poor model fit, we computed the Variance Inflation Factor[8] (VIF) score of each model. In particular, this was done in order to rule out any correlation between verbal and visual SORT or between OPT and MSIT scores. In none of the models was the VIF associated with any of the covariates at a level higher

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

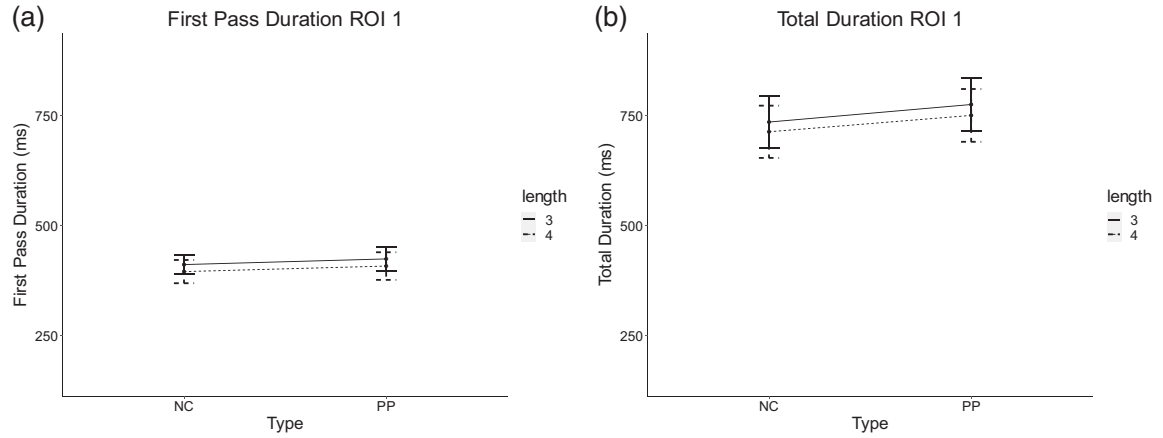


Figure 8. Experiment 2: Reading measures in Region of Interest 1. Error bars represent 95% confidence intervals.

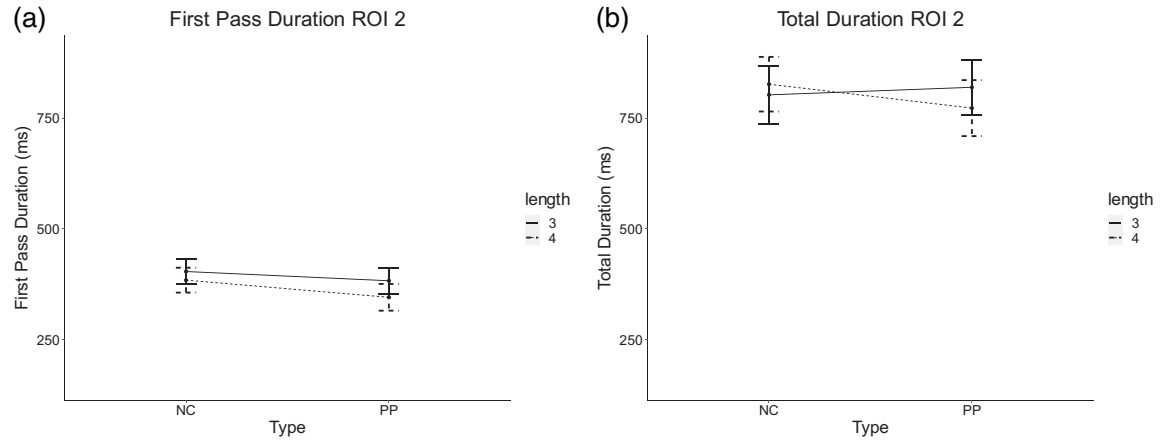


Figure 9. Experiment 2: Reading measures in Region of Interest 2. Error bars represent 95% confidence intervals.

than 2, indicating that, for each model, none of the variables could be reliably predicted based on the other model variables.

Discussion

In Experiment 2, we further explored the differences predicted by the UID between the processing of NCs and that of PPs with improved items. We expected longer reading times for all measures in the segments following the critical region for sentences containing NCs and sentences containing a longer critical structure. As in Experiment 1, we also expected these differences to be modulated by individual differences, such as English proficiency and WM abilities.

Our results partially replicate those of Experiment 1. We found no effects of length or type for FPD in ROI1, nor for TD in ROI1 and ROI2, nor for proficiency in any measure analyzed. In addition, we did find significant effects of length and type for Reg2CR in the same direction as Experiment 1 (with longer structures and PPs leading to more regressions).

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

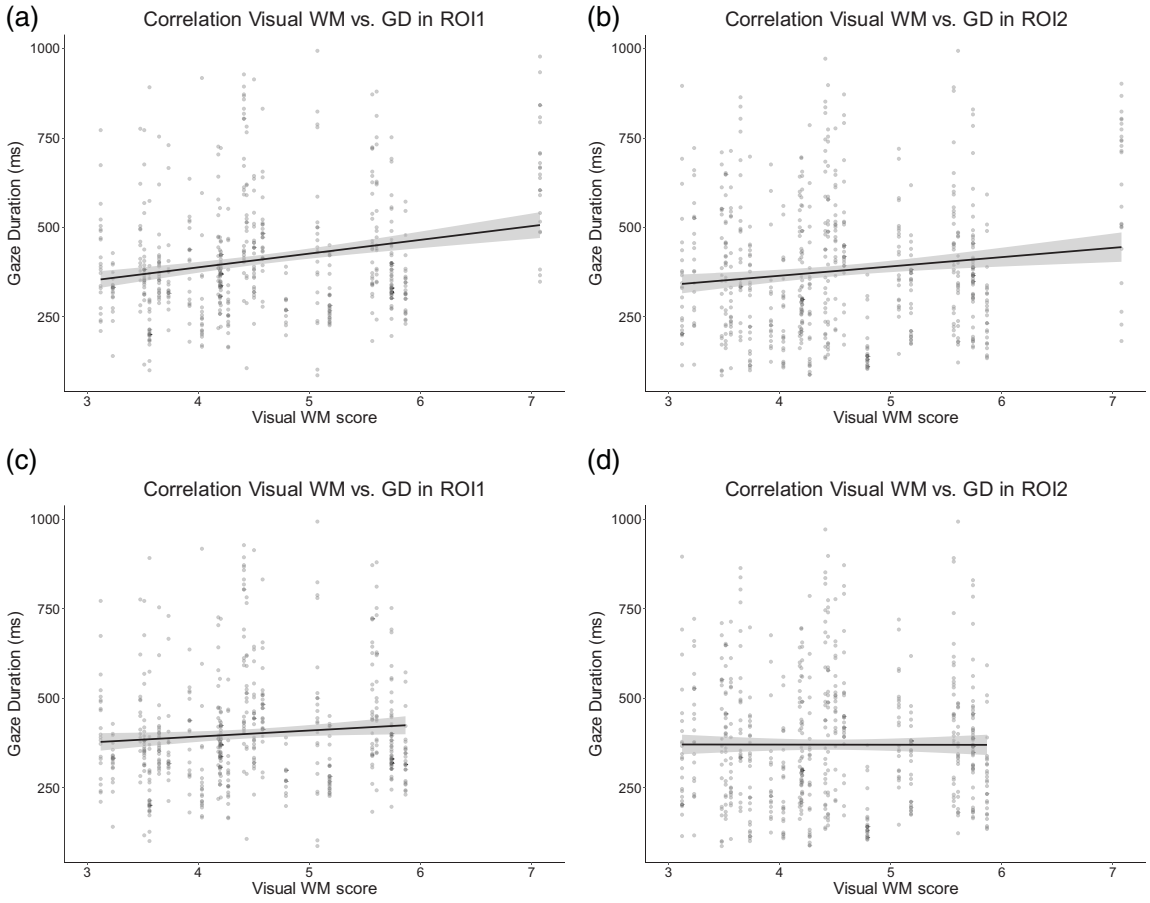


Figure 10. Experiment 2: Correlation between participants' visual WM scores and First Pass Duration in ROI1 (a) and ROI2 (b), found to be significant predictors in our model, and the same results (c and d, respectively) after removing the single outlier with very high Visual WM score. Gray shading indicates 95% confidence intervals for the regression line.

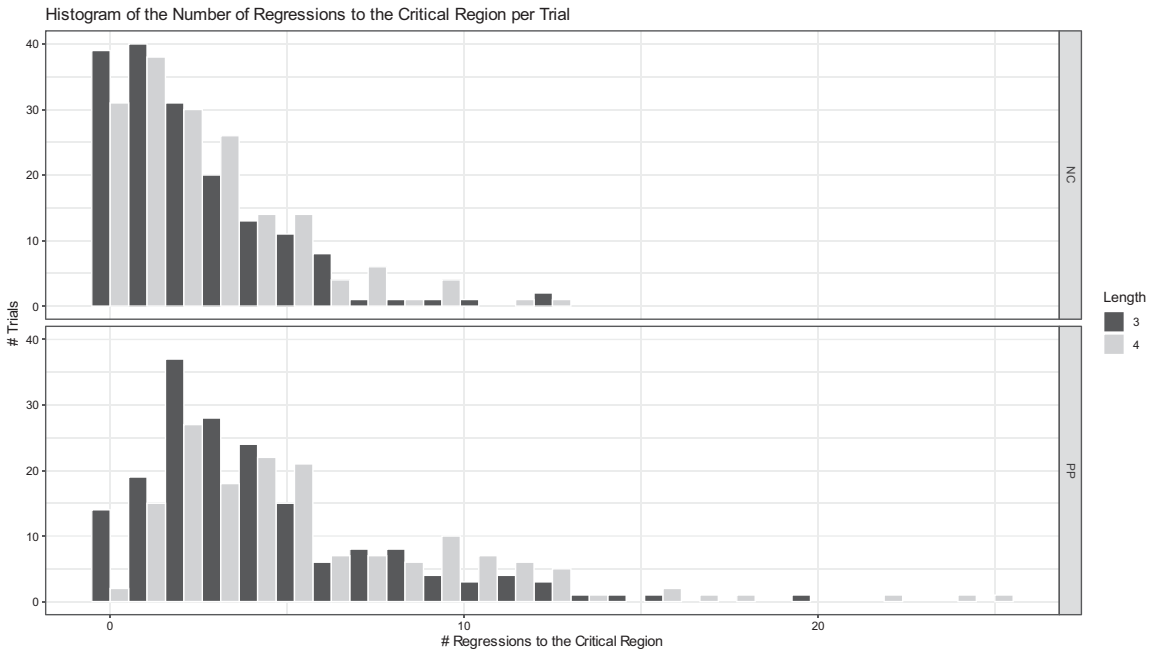


Figure 11. Experiment 2: Regressions onto the critical region.

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

Table 7. An overview of the results of Experiments 1 and 2. Effects that were no longer significant after outlier removal are not shown

	Experiment 1	Predicted	Experiment 2	Predicted
Critical	Reg2CR: PP > NC	No	Reg2CR: PP > NC	No
Region	Reg2CR: 6 > 4	Yes	Reg2CR: 4 > 3	Yes
ROI 1	↑DST → ↓TD	Yes		
ROI 2	↑DST → ↓TD	Yes	FPD: NC > PP	Yes

In contrast with Experiment 1, however, we found no effect of WM scores. This may reflect the fact that we used a different task in Experiment 2, and it is not clear whether we would have found a significant effect had we used a DST. More interestingly, we did find an effect of type on FPD in ROI2 in the predicted direction, namely, NCs led to longer FPDs than PPs. This effect constitutes evidence in favor of the UID, suggesting that NCs *are* harder to process than PPs.

General discussion

In two experiments, we investigated the online processing of nominal compounds (NC) compared to that of nouns modified by prepositional phrases (PP), which use more words and express roughly the same meaning. We assumed that NCs transfer more information per symbol, leading to a peak in information density. Based on the UID hypothesis, we predicted that these peaks would cause processing difficulty. We also assumed that the UID hypothesis would subsume all other sources of NC processing difficulty that have been both theorized and shown in the NC literature. In our online studies, we expected it to be reflected by more frequent regressions toward the NC and longer reading times in the subsequent text segments (as compared to PPs). The pattern of results found in the two experiments is summarized in Table 7.

Relevance of the findings to the UID

The clearest indication we have in favor of our predictions was the longer First Pass Durations (FPDs) in ROI2 for NCs in Experiment 2: As predicted, the denser structures evoked more processing difficulty than the less dense ones. This effect only became apparent after controlling for the shortcomings of Experiment 1. It was an early effect, only found in a segment further away from the critical segment (ROI2). In addition, we found a clear effect of length in the analysis of regressions toward the critical region (Reg2CR): Longer structures led to more Reg2CR in both studies. We interpret these as evidence in favor of the UID hypothesis, since longer structures are presumed to be associated with higher information transmission rate.[9]

Contrary to expectation, trials containing PPs had significantly more regressions than trials containing NCs. We believe this may be an artifact of two characteristics of this study: The position of the head noun and the large length difference between the structures. First, the head noun in PPs is the first word encountered by the

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

reader, thus leading to a long gap between the critical region head noun and the ROI1 tensed verb. When the reader arrives at ROI1, they need to regress in order to verify the agreement between the two words. In NC sentences, the head noun is the last word of the critical region, the reader probably still has it in working memory (WM) when they reach the tensed verb in ROI1, and hence no regression is necessary. Second, recall that a fixation is a Reg2CR if two conditions are met: The fixation is a regression (moved backward with respect to the previous fixation) and is positioned on the critical region. We chose to include regressions *inside* the critical region because of the region's length: Participants could experience difficulty inside the critical region itself, which would be hidden if we only considered regressions coming from subsequent text areas. But this decision meant that participants have many more chances to regress inside a PP than inside an NC, potentially leading to a spurious effect of Length in Reg2CR.

Interestingly, none of the measures was predicted by English proficiency in either experiment. Of course, participants were L1 speakers and therefore generally showed high proficiency, resulting in small variance in proficiency scores, which may have obscured the effects of proficiency on the measures. Note, however, that our findings are consistent with those of recent studies on the association between proficiency and prediction (Dijkgraaf et al., 2017; Kim & Grüter, 2021; Ito et al., 2018; Mitsugi, 2020). The integration difficulty expected in our study is linked to the notion of prediction, that is, the idea that the surrounding context influences the state of the language parsing system, which is then used to infer the upcoming signal (Kuperberg & Jaeger, 2016). Under this framework, in a way very much compatible with the UID formulation, one could attribute the parsing difficulty experienced upon encountering an NC/PP to a prediction error, where participants incorrectly expected a different sequence of upcoming words. Even though it is typically assumed that better proficiency necessarily leads to better prediction, our results are in line with recent studies suggesting that these two abilities are not as clearly associated – neither for L1 (Dijkgraaf et al., 2017) nor for L2 (Ito et al., 2018; Kim & Grüter, 2021; Mitsugi, 2020) – as has been typically assumed (see review in Kaan & Grüter, 2021).

Conversely, the digit span test (DST) score was a good predictor of total durations (TDs) in Experiment 1 and did present a negative (but not significantly different from zero) correlation with FPDs. Replacing the DST with serial order reconstruction tasks (SORTs) did not produce the expected results: Neither verbal nor visual SORT significantly predicted the duration measures. It is not clear how DST scores relate to SORT results and it is therefore not possible to say whether the results would be the same had we used a DST in Experiment 2. Further experiments are needed in order to better understand the relationship between the different WM measures and their impact on eye-tracking reading measures.

Given the abundance of literature showing that the UID holds for production, it may seem odd that we did not find clear results for comprehension. How can this be? In order to better understand these results, there are at least three different perspectives through which this data should be considered, which are discussed separately below.

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

Technical matters related to the data

One potential confounding factor in our results is technical matters related to the data. Because the sentences used in Experiments 1 and 2 were long, stimulus presentation was performed in two text lines, with the critical segment positioned at the end of the first line, and ROIs 1 and 2 positioned at the beginning of the second line. This caused three types of distortion in the data. First, in order to move from the critical segment toward ROI1, participants needed to perform a long saccade with a high chance of landing by mistake in ROI2. By definition, when a participant fixated in ROI2 *before* fixating in ROI1, then FPD in ROI1 was 0 ms. This led the trial to be discarded during trimming from the FPD ROI1 analysis, partially explaining the high number of trimmed trials in ROI1 (see Tables 3 and 6). Second, participants varied substantially the vertical position of their fixations while reading. For example, while reading *United States* in Fig. 3, some participants performed fixations too low, closer to the second text line than the first one. When extracting the reading measures, this was calculated as equivalent to "skipping" ROI1, again causing FPD to be 0 ms. Third, regressions were not easily identifiable as left-saccades (e.g., moving from ROI1 to the critical region required a right-saccade). Thus, we had to redefine "regression" based on the previous fixation's interest area: If the previous fixation was in the interest area of a word that followed the current fixation's word, then the current fixation counted as a regression. The disadvantage of this definition is that it disregarded regressions coming from outside any interest areas. In order to alleviate the impact of these shortcomings, we took them into account when implementing our correction algorithm (see Supplementary Materials). Our subjective review of the corrected data concluded that the correction did substantially improve its quality. However, we suggest that future studies be run with stimulus presentation in a single text line in order to avoid the aforementioned problems.

In addition, given the exploratory nature of the study, we chose to investigate the time course of the predicted difficulty by only analyzing two reading measures (FPD and TD), thus avoiding too many comparisons. Given the results reported here, it may be that our choice of reading measures was unlucky and that other measures might have provided clearer evidence favoring the UID, given that eye-tracking reading measures have been found not to correlate much with one another (see von der Malsburg & Angele, 2017). Furthermore, our results may also reflect the indirect nature of our paradigm, which depends on participants' difficulty spilling over to the subsequent text segments. While both options are in principle possible, it is important to note that the kind of spillover-based analysis we chose is well established in the literature, and numerous studies on the processing of other structures have found significant differences in spillover regions analyzed the way we did using the exact same reading measures we chose (e.g., Christianson et al., 2017; Jared & O'Donnell, 2017; Paape & Vasishth, 2022; Pickering & Traxler, 1998).

Cognitive load differences in comprehension vs. production

When interpreting the data reported here, one should also consider the differences between production and comprehension, and in particular the differences in

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

cognitive demands associated with each task type. The relationship between production and comprehension is an open area of research, not yet clearly understood, and results that hold for one task type sometimes do not hold for the other (for a review, see Iraola Azpiroz et al., 2019a, 2019b). As discussed earlier, most other studies favoring the UID have probed production. Perhaps real-time production is more difficult, involving turn-taking, linguistic parsing, motor coordination, and typically relying on real-world information, making speakers more likely to be affected by peaks and troughs of information density, as the associated cognitive costs may harm their communicative goals. Conversely, reading is static, and the communicative goals of the writer are presumably not the same as those of a conversation partner. Thus, the reader may not be taxed to the same extent as a speaker, the effect may be subtler, and experiments investigating the UID in reading may actually require starker peaks in information density to unlock the same effects. Hence, maybe it is possible to find results favoring the UID by imposing an additional task that taxes participant WM while reading to exacerbate information density peak-related costs. This would explain the findings of Sikos et al., 2017 in support of the UID, since two aspects of their design provided additional cognitive load. First, they used a G-Maze task that forced participants to make decisions about the sentence as they went through it. Second, their experiment incorporated a prediction aspect by manipulating the subject (*the journalist published/the man evaluated the carefully written essay*). Kuperberg and Jaeger (2016) point out that prediction is costly and suggest that speakers balance this cost with the reliability of their predictions and with how useful their predictions are in advancing their communicative goal.

Our study illustrates the importance of converging evidence from different domains: Despite the evidence for production, the UID hypothesis may not hold as clearly when considering well-designed reading studies, for which it may actually be harder to find evidence in its favor. Of course, this study is one of the first to focus on the UID while reading, and much more exploration is necessary to determine its validity.

The role of experience with NC use

Finally, it is important to consider the relevance of previous participant experience with NCs to our results. When preparing the experiments reported here, we had good reason to presume that NCs are not straightforward to process: NCs had been thus theorized in the literature, and a number of behavioral studies had shown evidence favoring this assumption. However, the few significant differences reported here cast some doubt on this presumption. Are NCs *really* denser than PPs?

In order to answer this question, we used a language model to estimate the informational content of our items. The language model we used was OpenAI's GPT-2 (Radford et al., 2019), a pre-trained version of which can be downloaded from the web. This ready-to-use model was trained on a dataset scraped from millions of links referred to by Reddit users, contemplating pages of all sorts of domains, including commercial pages (e.g., eBay, Apple, Craig's list), journalistic pages (New York Times, BBC, Reuters), wikis (Wikipedia, Wikia, Gamepedia), other user-created content

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

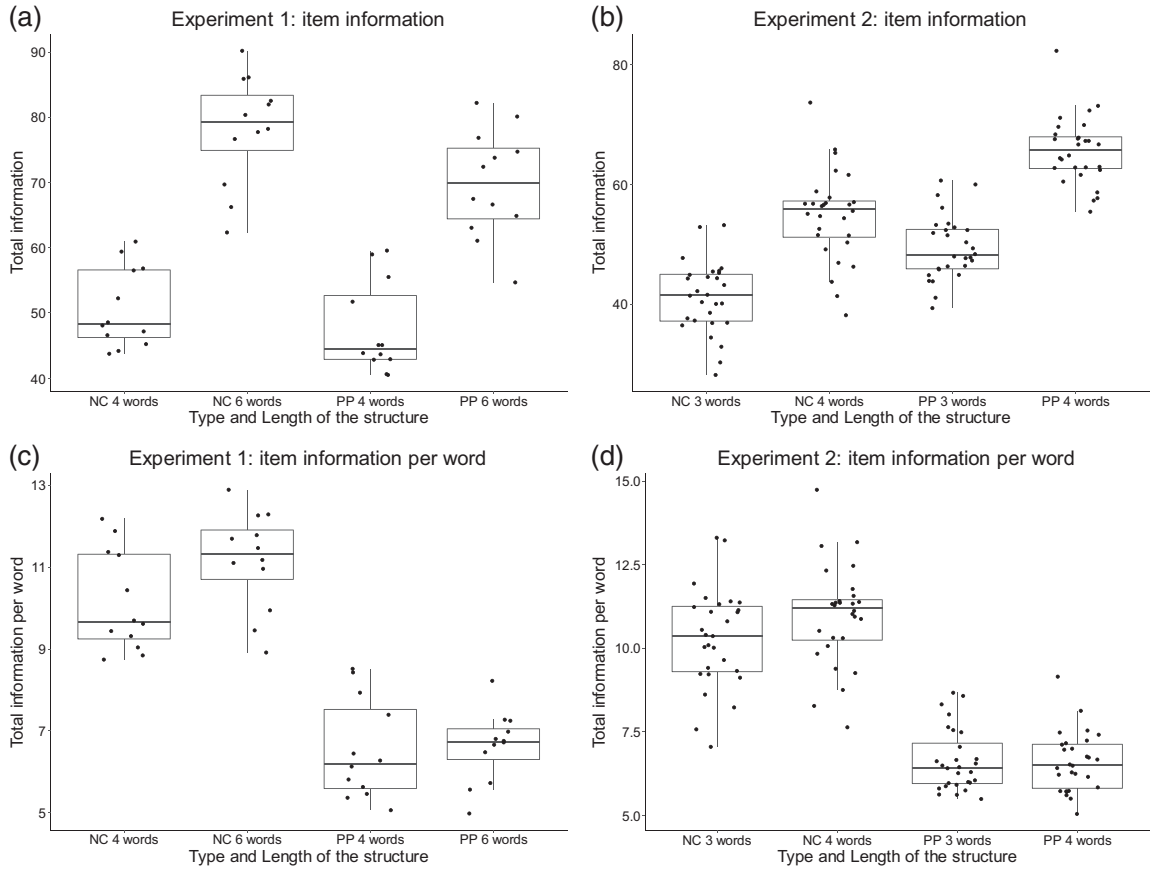


Figure 12. The information of items in Experiment 1 and 2. Graphs show the total amount (a and b) and the information per word (c and d). In all graphs, each point represents an item.

(Urban Dictionary, Pastebin, Medium, Stack Exchange), as well as academic pages (Nasa, Stanford, NIH). Given a context c (a set of words, e.g., "pharmaceutical market") and a continuation word w (e.g., "size"), the model produces probabilities $P(w|c)$. The informational content of the word w is calculated as $-\log(P(w|c))$. Of course, to calculate the information of a subsequent word w_2 (e.g., "increase" after "size") we use w in its context, that is, we calculate $-\log(P(w_2|c,w)$. Since information is additive, the information content of both "size" and "increase" is the sum of their individual contents (see Appendix E for details).

We used the language model and the aforementioned procedure to calculate the information content of all of our items. The results of the model can be seen in Fig. 12. Considered generally, the NCs used in our study do carry more information per word than their PP counterparts. While the total NC information amount is not very different from that of PPs, the distributions become much more clearly divided when considered the amount of information *per word*.[10]

Why, then, do we not find clear processing difficulties caused by NCs relative to PPs? There are two factors that may have affected the study results. First, a large portion of our participants were university students, who might be used to reading scientific articles, and thus NCs. Presumably, their internal parser already associates a higher probability (and thus a lower density) to NCs.

Second, NC use in certain registers is changing. Consider Biber and Gray's (2011) diachronic study of NCs in several English registers. Even though they found an

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

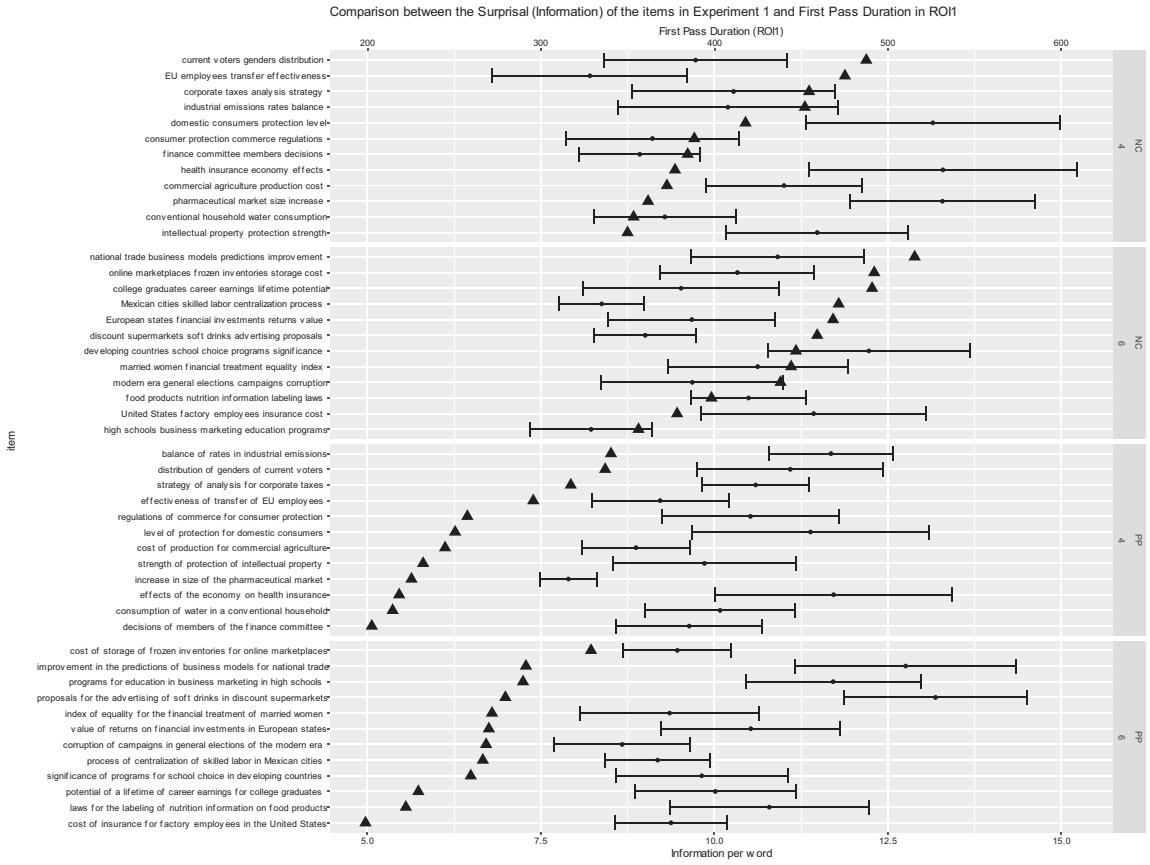


Figure 13. Figure shows items of Experiment 1 ordered by their surprisal per word, as indicated by the triangles, and separated by condition. Mean and standard error of FPD in ROI1 are indicated by the dot and the error bars. Similar comparisons between reading measures and *total* surprisal of each item are also available in the Supplementary Materials.

increase in NC use in academic and journalistic texts (see Fig. 2), this was not the case for the other registers they investigated (novels and drama), suggesting that NC probability may still be considerably low in general English and corroborating the output of the language model. However, for the academic register specifically, although their latest reported data are from 2005, we have no reason to think that the trend has stopped there: The probability associated with the structure is increasing, making them less informational.

The information formula discussed above is also known as *surprisal* in the literature and is at the core of Surprisal Theory, a theory hypothesizing that a word's processing difficulty is proportional to its surprisal (Levy, 2008): More informative (surprising) words should lead to more processing difficulty. Given the calculated information values above, it may be worth asking whether Surprisal Theory would have fared better at predicting the processing difficulty experienced by participants.[11] Considering the aforementioned factors affecting the results, we would expect surprisal (information) to be a bad predictor of processing difficulty. Unfortunately, we do not have enough data to add item information values to our models. However, Fig. 13 shows the information content of each item of Experiment 1 along with the observed FPD values in ROI1. As expected, there does not seem to be a clear relationship between FPD values and information.[12] Further research should address this question in a more careful way.

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

Overall, this highlights the role of experience in language processing in general and in particular in the processing of complex structures. Even though NCs are *generally* dense, they may not be so for our specific combination of participants and items.

Conclusion

This study provides evidence in favor of the UID hypothesis through the investigation of a structure typically considered hard to process, namely complex nominal compounds consisting of three or more words. In one of the first studies to investigate the UID hypothesis from the point of view of comprehension, we compared NCs with a much longer structure (PPs) that spreads the information conveyed more evenly through time. The results reported in Experiments 1 and 2 indicate that NCs do lead to comprehension difficulty, although the pattern of results was not as clear as predicted.

The results reported here have two key implications for the UID. First, we interpret our results as reflecting differences between the processes of production and comprehension, namely the fact that production tends to be harder than comprehension. As a result, it may be beneficial in future reading studies to include an additional parallel task in order to increase participants' cognitive load and this way evoke the effects predicted by the UID hypothesis. Second, these results highlight the importance of considering the role of experience in the design of UID studies. NC use is becoming more common in recent years, especially among the academic population that is typically recruited in psycholinguistic studies, and we argue that in our case their experience with the structure led to less difficulty in its processing.

In sum, this study gives rise to two key areas that should form an integral part of our understanding of the UID, namely how the comprehension-production asymmetry and experience with a given register impact the way the UID hypothesis has been operationalized in the literature.

Replication package. All research materials, data, and analysis code are available at: https://osf.io/tqspf.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0142716424000092

Acknowledgements. We are grateful for the help of many people in bringing this project to fruition in the present paper. Numerous students and colleagues provided assistance in developing materials, testing participants, piloting earlier versions of the study, and discussing the results: Reza Akhtar, Anthony Akinbodunse, Ian Brenckle, Lindsay Coffin, Liz Dovenberg, Mary Elliot, Neiloufar Family, Maialen Iraola Azpiroz, Gunnar Jacob, Abdullah Jelelati, Alice Johnson, Kalliopi Katsika, Maria Klatte, Victor Kuperman, Tenyse Wells, Nariman Utegaliyev, Fransisca Hapsari, Jamie Nisbet, Lisa Martinek, Liberus Ogbonna Ogochukwu, and Hannah Powers. Lianna Fortune and Juhani Järvikivi generously collected the data for Experiment 1 at the University of Alberta. Audiences at several conferences and university colloquia provided feedback on the ideas presented here, and the Kaiserslautern Scientific Writing Group provided helpful comments on previous versions of this article. This work was funded by the Rheinland-Pfälzische Technische Universität via a doctoral fellowship to the first author through the Rhineland-Palatinate State Research Initiative and via a faculty start-up grant to the third author.

Notes

1 NCs have been treated inconsistently in the literature, with studies varying widely on their terminology and on their definition. In particular, they often vary on what they allow in premodifying position (e.g., whether they allow adjectives or only nouns). We ignore these differences in our discussion of the literature.
2 All research materials, data, and analysis code are available at https://osf.io/tqspf
3 Note that our definition of First Pass Duration is slightly different from that used in other studies (e.g., Cook & Wei, 2019; Schaeffer et al., 2019). The measure we use has been referred to by names such as *right-bounded reading time* (Gordon et al., 2006) or *quasi-first-pass time* (Traxler et al., 2002). Contrasting with the rest of the literature, it is also called *gaze duration* by the script we used to extract it (Dan, 2020).
4 All models used the formula DV ~ OPT + DST + length*type + (1 + length * type | subject) + (1 + DST + OPT + type*length | item).
5 However, note that the information content of a sentence may also be a way to explain garden path effects (see e.g., Hale, 2001; Levy, 2008). By this explanation, as the participant reads the sentence, they produce a partial parse along with a probability distribution of the likely content to appear next. A garden path effect would happen when the probability assigned to the content that is *observed* next is very low. From the point of view of information theory, this low probability would be translated into a very high amount of information transmitted by that content (say, a word). If this amount of information is higher than the channel capacity, the UID hypothesis would predict difficulty.
6 Here, we use the term collocation broadly, meaning any "familiar recurrent expression" (Gledhill, 2000, p. 6) commonly found in the English language and easily identifiable by native speakers, such as "strong coffee" or "social media".
7 All models used the formula DV ~ OPT + SpellingScore + length*type + VerbalWM + VisualWM + (1 + length * type | subject) + (1 + VerbalWM + VisualWM + type*length | item)
8 The Variance Inflation Factor is a measure of the reliability with which one of the variables in the model could be estimated based on the other variables of the model. As a rule of thumb, VIFs higher than 10 are problematic (see, e.g., Craney & Surles, 2002). We used R's *car* package (Fox & Weisberg, 2019) in order to compute these values.
9 However, this latter effect should be considered with care. One reviewer suggested that this effect, as well as the effect of phrase type discussed in the subsequent paragraph, could be explained by the difference between the lengths of the critical region in the two types of structures: longer structures and PPs would lead to more regressions simply because readers have a higher chance of landing on them upon performing saccades. In order to take the length difference into account, we fit two new models normalizing the regression count by the length of the critical region. That is, the new models predict instead the rate $\frac{\text{Number of regressions to the critical region}}{\text{Number of characters of the critical region}}$. In R, this is done by simply adding the term offset(log(length_in_characters)) to the Reg2CR model formulas.

Indeed, in the new models the effect of length vanishes, (see Tables C6 and D6), demonstrating the importance of controlling for length in future experiments. In addition, the effect of phrase type becomes non-significant in Experiment 1, but remains significant in Experiment 2, where PPs still led to significantly more regressions than NCs. We discuss possible reasons for this effect in the next paragraph.
10 Note that there is no special reason why we calculate density by dividing the total amount of information by the number of words in the structures. However, this is a common choice in the literature (see Meister et al., 2021, for a discussion).
11 We thank a reviewer for this suggestion.
12 Similar graphs were made for FPD and TD, for ROI1 and ROI2, for both experiments. However, they look substantially similar to Figure 13 and are not reported here (see Supplementary Materials).

References

Allan, D. (2006). *Oxford placement test 1: Audio CD*. Oxford, UK: Oxford University Press.

Andrews, S., Veldre, A., & Clarke, I. E. (2020). Measuring lexical quality: The role of spelling ability. *Behavior Research Methods*, **52**(6), 2257–2282. https://doi.org/10.3758/s13428-020-01387-3

Baddeley, A. (2011). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, **63**, 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, **68**(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bartolic, L. (1978). Nominal compounds in technical English. In L. Trimble, M. Trimble, & K. Drobnic (Eds.), *English for specific purposes: Science and technology* (pp. 257–277). Corvallis, OR: Oregon State University.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**(1), 1–48. (R package version 1.1-18-1). https://doi.org/10.18637/jss.v067.i01

Bauer, L., & Tarasova, E. (2013). The meaning link in nominal compounds. *SKASE Journal of Theoretical Linguistics*, **10**(3), 2–18. Retrieved from http://www.skase.sk/Volumes/JTL24/pdf_doc/01.pdf

Berg, T. (2016). The semantic structure of English and German compounds: Same or different? *Studia Neophilologica*, **88**(2), 148–164. https://doi.org/10.1080/00393274.2015.1135758

Bhatia, V. K. (1992). Pragmatics of the use of nominals in academic and professional genres. In L. F. Bouton & Y. Kachru (Eds.), *Pragmatics and language learning: Monograph series* (Vol. **3**, pp. 217–230). Urbana, IL, USA: University of Illinois. Retrieved from https://eric.ed.gov/?id = ED395531

Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, **9**(1), 2–20. https://doi.org/10.1016/j.jeap.2010.01.001

Biber, D., & Gray, B. (2011). Grammatical change in the noun phrase: The influence of written language use. *English Language & Linguistics*, **15**(2), 223–250. https://doi.org/10.1017/S1360674311000025

Blignaut, P., Holmqvist, K., Nyström, M., & Dewhurst, R. (2014). Improving the accuracy of video-based eye tracking in real time through post-calibration regression. In M. Horsley, M. Eliot, B. A. Knight, & R. Reilly (Eds.), *Current trends in eye tracking research* (pp. 77–100). Cham: Springer. https://doi.org/10.1007/978-3-319-02868-2_5

Brekle, H. E. (1976). *Generative Satzsemantik im System der englischen Nominalkomposition [Generative sentential semantics in the English nominal system]*. Munich: Fink.

Carrió Pastor, M. L. (2008). English complex noun phrase interpretation by Spanish learners. *Revista Española de Lingüística Aplicada*, **21**, 27–44. Retrieved from https://dialnet.unirioja.es/servlet/articulo?codigo = 2925910

Carrió Pastor, M. L., & Candel Mora, M. Á. (2013). Variation in the translation patterns of English complex noun phrases into Spanish in a specific domain. *Languages in Contrast*, **13**(1), 28–45. https://doi.org/10.1075/lic.13.1.02car

Christianson, K., Luke, S. G., Hussey, E. K., & Wochna, K. L. (2017). Why reread? Evidence from garden-path and local coherence structures. *The Quarterly Journal of Experimental Psychology*, **70**(7), 1380–1405. https://doi.org/10.1080/17470218.2016.1186200

Collins, M. X. (2014). Information density and dependency length as complementary cognitive models. *Journal of Psycholinguistic Research*, **43**, 651–681. https://doi.org/10.1007/s10936-013-9273-3

Cook, A. E., & Wei, W. (2019). What can eye movements tell us about higher level comprehension? *Vision*, **3**(3), 45.

Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering*, **14**(3), 391–403. https://doi.org/10.1081/QEN-120001878

Dan. (2020). *Get Reading Measures* [Online forum post]. Retrieved 21 June 2021 from https://www.sr-research.com/support/showthread.php?tid = 26

de Almeida, R. G., & Libben, G. (2005). Changing morphological structures: The effect of sentence context on the interpretation of structurally ambiguous English trimorphemic words. *Language and Cognitive Processes*, **20**, 373–394. https://doi.org/10.1080/01690960444000232

Dijkgraaf, A., Hartsuiker, R. J., & Duyck, W. (2017). Predicting upcoming information in native-language and non-native-language auditory word recognition. *Bilingualism: Language and Cognition*, **20**(5), 917–930. https://doi.org/10.1017/S1366728916000547

Downing, P. (1977). On the creation and use of English compound nouns. *Language*, **53**(4), 810–842. https://doi.org/10.2307/412913

Estes, Z., & Jones, L. L. (2006). Priming via relational similarity: A copper horse is faster when seen through a glass eye. *Journal of Memory and Language*, **55**(1), 89–101. https://doi.org/10.1016/j.jml.2006.01.004

Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, **40**(4), 296–340. https://doi.org/10.1006/cogp.1999.0730

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Thousand Oaks, CA: Sage.

Frank, A. F., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In Proceedings of the annual meeting of the Cognitive Science Society (Vol. 30). Retrieved from https://escholarship.org/uc/item/7d08h6j4

Frank, S. L., Monaghan, P., & Tsoukala, C. (2019). Neural network models of language acquisition and processing. In P. Hagoort (Ed.), *Human language: From genes and brain to behavior* (pp. 277–293). Cambridge, MA: MIT Press. Retrieved from https://www.mpi.nl/publications/item_3347596

Gagné, C. L., & Shoben, E. J. (1997). Influence of thematic relations on the comprehension of modifier–noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **23**(1), 71–87. https://doi.org/10.1037/0278-7393.23.1.71

Gagné, C. L., & Spalding, T. L. (2009). Constituent integration during the processing of compound words: Does it involve the use of relational structures? *Journal of Memory and Language*, **60**(1), 20–35. https://doi.org/10.1016/j.jml.2008.07.003

Geer, S. E., Gleitman, H., & Gleitman, L. (1972). Paraphrasing and remembering compound words. *Journal of Verbal Learning and Verbal Behavior*, **11**(3), 348–355. https://doi.org/10.1016/S0022-5371(72)80097-5

Gibson, E. (2001). The dependency locality theory: A distance-based theory of linguistic complexity. In A. P. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain: Papers from the first mind articulation project symposium* (pp. 95–126). Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/3654.003.0008

Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, **23**(5), 389–407. https://doi.org/10.1016/j.tics.2019.02.003

Gledhill, C. J. (2000). *Collocations in science writing*. Tübingen: Gunter Narr Verlag.

Gleitman, L. R., & Gleitman, H. (1970). *Phrase and paraphrase: Some innovative uses of language*. New York: Norton.

Gordon, P. C., Hendrick, R., Johnson, M., & Lee, Y. (2006). Similarity-based interference during language comprehension: Evidence from eye tracking during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **32**(6), 1304.

Granville Hatcher, A. (1960). An introduction to the analysis of English noun compounds. *Word*, **16**(3), 356–373. https://doi.org/10.1080/00437956.1960.11659738

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). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.3115/1073336.1073357

Hornof, A. J., & Halverson, T. (2002). Cleaning up systematic error in eye-tracking data by using required fixation locations. *Behavior Research Methods, Instruments, & Computers*, **34**(4), 592–604. https://doi.org/10.3758/BF03195487

Horsella, M., & Pérez, F. (1991). Nominal compounds in chemical English literature: Toward an approach to text typology. *English for Specific Purposes*, **10**(2), 125–138. https://doi.org/10.1016/0889-4906(91)90005-H

Inhoff, A. W., Radach, R., & Heller, D. (2000). Complex compounds in German: Interword spaces facilitate segmentation but hinder assignment of meaning. *Journal of Memory and Language*, **42**(1), 23–50. https://doi.org/10.1006/jmla.1999.2666

Iraola Azpiroz, M., Allen, S. E. M., Katsika, K., & Fernandez, L. B. (2019a). Psycholinguistic approaches to production and comprehension in bilingual adults and children. *Linguistic Approaches to Bilingualism*, **9**(4/5), 505–513. https://doi.org/10.1075/bct.117.01azp

Iraola Azpiroz, M., Allen, S. E. M., Katsika, K., & Fernandez, L. B. (Eds.). (2019b). Special issue of Linguistic Approaches to Bilingualism: Psycholinguistic approaches to production and comprehension in bilingual adults and children. *Linguistic Approaches to Bilingualism*, **9**(4/5). https://doi.org/10.1075/bct.117

Ito, A., Corley, M., & Pickering, M. J. (2018). A cognitive load delays predictive eye movements similarly during L1 and L2 comprehension. *Bilingualism: Language and Cognition*, **21**(2), 251–264. https://doi.org/10.1017/S1366728917000050

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, **61**(1), 23–62. https://doi.org/10.1016/j.cogpsych.2010.02.002

Jared, D., & O'Donnell, K. (2017). Skilled adult readers activate the meanings of high-frequency words using phonology: Evidence from eye tracking. *Memory & Cognition*, **45**, 334–346. https://doi.org/10.3758/s13421-016-0661-4

Jones, D., Farrand, P., Stuart, G., & Morris, N. (1995). Functional equivalence of verbal and spatial information in serial short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **21**(4), 1008–1018. https://psycnet.apa.org/doi/10.1037/0278-7393.21.4.1008

Kaan, E., & Grüter, T. (2021). Prediction in second language processing and learning: Advances and directions. In E. Kaan & T. Grüter (Eds.), *Prediction in second language processing and learning* (pp. 1–24). Amsterdam: John Benjamins. https://doi.org/10.1075/bpa.12.01kaa

Kim, H., & Grüter, T. (2021). Predictive processing of implicit causality in a second language: A visual-world eye-tracking study. *Studies in Second Language Acquisition*, **43**(1), 133–154. https://doi.org/10.1017/S0272263120000443

Krott, A., Libben, G., Jarema, G., Dressler, W., Schreuder, R., & Baayen, H. (2004). Probability in the grammar of German and Dutch: Interfixation in triconstituent compounds. *Language and Speech*, **47**(1), 83–106. https://doi.org/10.1177/00238309040470010401

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, **31**(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299

Kuperman, V., Bertram, R., & Baayen, R. H. (2008). Morphological dynamics in compound processing. *Language and Cognitive Processes*, **23**(7), 1089–1132. https://doi.org/10.1080/01690960802193688

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H., et al. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, **82**(13), 1–26. (R package version 3.0.1) https://doi.org/10.18637/jss.v082.i13

Kvam, A. M. (1990). Three-part noun combinations in English, composition – meaning – stress. *English Studies: A Journal of English Language and Literature*, **71**(2), 152–161. https://doi.org/10.1080/00138389008598684

Lees, R. B. (1960). *The grammar of English nominalizations*. Bloomington: Indiana University Press.

Lees, R. B. (1970). Problems in the grammatical analysis of English nominal compounds. In M. Bierwisch & K. E. Heidolph (Eds.), Progress in linguistics (pp. 174–186). The Hague: Mouton. https://doi.org/10.1515/9783111350219.174

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, **44**(2), 325–343. https://doi.org/10.3758/s13428-011-0146-0

Levi, J. N. (1973). Where do all those other adjectives come from? In C. Corum, T. C. Smith-Stark, & A. Weiser (Eds.), *Papers from the 9th regional meeting of the Chicago Linguistic Society* (pp. 332–345). Retrieved from https://www.ingentaconnect.com/contentone/cls/pcls/1973/00000009/00000001/art00030

Levi, J. N. (1978). *The syntax and semantics of complex nominals*. New York: Academic Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, **106**(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Levy, R., & Jaeger, T. F. (2006). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. C. Platt, & T. Hoffman (Eds.), *Proceedings of the 19th International Conference on Neural Information Processing Systems* (pp. 849–856). Cambridge, MA: MIT Press. Retrieved from https://proceedings.neurips.cc/paper/2006/hash/c6a01432c8138d46ba39957a8250e027-Abstract.html

Limaye, M., & Pompian, R. (1991). Brevity versus clarity: The comprehensibility of nominal compounds in business and technical prose. *The Journal of Business Communication*, **28**(1), 7–21. https://doi.org/10.1177/002194369102800102

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, **6**, 1171. https://doi.org/10.3389/fpsyg.2015.01171

Lopopolo, A., Frank, S. L., van den Bosch, A., & Willems, R. M. (2017). Using stochastic language models (SLM) to map lexical, syntactic, and phonological information processing in the brain. *PLOS One*, **12**(5), e0177794. https://doi.org/10.1371/journal.pone.0177794

Meister, C., Pimentel, T., Haller, P., Jäger, L., Cotterell, R., & Levy, R. (2021). Revisiting the Uniform Information Density hypothesis. In M.-F. Moens, X. Huang, L. Specia, & S. W. Tau Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 963–980).

Stroudsburg, PA: Association for Computational Linguistics. http://doi.org/10.18653/v1/2021.emnlp-main.74

Merkx, D., & Frank, S. L. (2021). Human sentence processing: Recurrence or attention? In E. Chersoni, N. Hollenstein, C. Jacobs, Y. Oseki, L. Prévot, & E. Santus (Eds.), *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 12–22). Online. https://doi.org/10.18653/v1/2021.cmcl-1.2

Mitsugi, S. (2020). Generating predictions based on semantic categories in a second language: A case of numeral classifiers in Japanese. *International Review of Applied Linguistics in Language Teaching*, **58**(3), 323–349. https://doi.org/10.1515/iral-2017-0118

Montero, B. (1996). Technical communication: Complex nominals used to express new concepts in scientific English-causes and ambiguity in meaning. *The ESPecialist*, **17**(1), 57–72. Retrieved from https://revistas.pucsp.br/index.php/esp/article/view/9476/7042

Olshtain, E. (1981). English nominal compounds and the ESL/EFL reader. In M. Hines & W. Rutherford (Eds.), *On TESOL '81: Selected papers from the fifteenth Annual Conference of Teachers of English to Speakers of other Languages* (pp. 153–168). Washington, DC: TESOL. Retrieved from https://eric.ed.gov/?id = ED223079

Paape, D., & Vasishth, S. (2022). Does conscious rereading lead to targeted regressions in garden-path sentences? Data from a novel stop-and-reread paradigm. *PsyArXiv preprint*. https://doi.org/10.31234/osf.io/vjyfz

Pérez Ruiz, L. (2006). Unravelling noun strings: Toward an approach to the description of complex noun phrases in technical writing. *ES: Revista de Filología Inglesa*, **27**(1), 163–174. Retrieved from https://dialnet.unirioja.es/servlet/articulo?codigo = 2210385

Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds.), *Precursors of functional literacy* (pp. 189–213). Amsterdam: Benjamins. https://doi.org/10.1075/swll.11.14per

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, **122**(3), 280–291. https://doi.org/10.1016/j.cognition.2011.10.004

Pickering, M. J., & Traxler, M. J. (1998). Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **24**(4), 940. https://doi.org/10.1037/0278-7393.24.4.940

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria. Retrieved from https://www.r-project.org/ (Version 3.4.0)

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*. Retrieved from https://github.com/openai/gpt-2

Salager, F. (1984). Compound nominal phrases in scientific-technical literature: Proportion and rationale. In A. K. Pugh & J. M. Ulijn (Eds.), *Reading for professional purposes: Studies in native and foreign languages* (pp. 136–145). London: Heinemann.

Schaeffer, M., Nitzke, J., Tardel, A., Oster, K., Gutermuth, S., & Hansen-Schirra, S. (2019). Eye-tracking revision processes of translation students and professional translators. *Perspectives*, **27**(4), 589–603. https://doi.org/10.1080/0907676X.2019.1597138

Schmidtke, D., Kuperman, V., Gagné, C. L., & Spalding, T. L. (2016). Competition between conceptual relations affects compound recognition: The role of entropy. *Psychonomic Bulletin & Review*, **23**(2), 556–570. https://doi.org/10.3758/s13423-015-0926-0

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, **27**(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sikos, L., Greenberg, C., Drenhaus, H., & Crocker, M. W. (2017). Information density of encodings: The role of syntactic variation in comprehension. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 3168–3173). Austin, TX: Cognitive Science Society.

Swales, J. (1974). *Writing scientific English*. London: Thomas Nelson and Sons.

Tobin, M. J. (2002). Compliance (COMmunicate PLease wIth less abbreviations, noun clusters, and exclusiveness). *American Journal of Respiratory and Critical Care Medicine*, **166**(12), 1534–1536. https://doi.org/10.1164/rccm.2211001

Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, **47**(1), 69–90. https://doi.org/10.1006/jmla.2001.2836

Trimble, L. (1985). *English for science and technology: A discourse approach*. Cambridge: Cambridge University Press.

Varantola, K. (1984). *On noun phrase structures in engineering English*. Turku: Turun Yliopisto.

Vasishth, S. (2010). Integration and prediction in head-final structures. In H. Yamashita, Y. Hirose, & J. Packard (Eds.), *Processing and producing head-final structures* (pp. 349–367). Dordrecht: Springer. https://doi.org/10.1007/978-90-481-9213-7_16

von der Malsburg, T., & Angele, B. (2017). False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language*, **94**, 119–133. https://doi.org/10.1016/j.jml.2016.10.003

Wambach, D., Lamar, M., Swenson, R., Penney, D. L., Kaplan, E., & Libon, D. J. (2011). Digit Span. In J. S. Kreutzer, J. DeLuca, & B. Caplan (Eds.), *Encyclopedia of clinical neuropsychology* (pp. 844–849). New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-79948-3_1288

Warren, B. (1978). *Semantic patterns of noun-noun compounds*. Gothenburg: Acta Universitatis Gothoburgensis.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & van den Bosch, A. (2016). Prediction during natural language comprehension. *Cerebral Cortex*, **26**(6), 2506–2516. https://doi.org/10.1093/cercor/bhv075

Williams, R. (1984). A cognitive approach to English nominal compounds. In A. K. Pugh & J. M. Ulijn (Eds.), *Reading for professional purposes: Studies in native and foreign languages* (pp. 136–145). London: Heinemann.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Zhang, Y., & Hornof, A. J. (2011). Mode-of-disparities error correction of eye-tracking data. *Behavior Research Methods*, **43**(3), 834–842. https://doi.org/10.3758/s13428-011-0073-0

Zhang, Y., & Hornof, A. J. (2014). Easy post-hoc spatial recalibration of eye tracking data. In *Proceedings of the symposium on eye tracking research and applications* (pp. 95–98). New York: Association for Computing Machinery. https://doi.org/10.1145/2578153.2578166

Appendix A Experiment 1 items

Table A1. List of all items with length 4

1	NC	In some cases, pharmaceutical market size increase is driven by the competition in Western countries.
	PP	In some cases, increase in size of the pharmaceutical market is driven by the competition in Western countries.
2	NC	For our investigation, conventional household water consumption is considered by our team in numerous ways.
	PP	For our investigation, consumption of water in a conventional household is considered by our team in numerous ways.
3	NC	In some cases, intellectual property protection strength is increased by a reform to national patents.
	PP	In some cases, the strength of protection of intellectual property is increased by this reform to national patents.
4	NC	In present times, the domestic consumer protection level is strengthened by the governments of wealthy countries.
	PP	In present times, the level of protection for domestic consumers is strengthened by the governments of wealthy countries.
5	NC	In current times, EU employee transfer effectiveness is debated by the experts in the foreign office.
	PP	In current times, effectiveness of transfer of EU employees is debated by the experts in the foreign office.
6	NC	In present times, health insurance economy effects are researched by the analysts of financial institutions.
	PP	In present times, effects of the economy on health insurance are researched by the analysts of financial institutions.
7	NC	In many countries, current voter gender distribution is evaluated by the Office of Voter Registration.
	PP	In many countries, the distribution of genders of current voters is evaluated by the Office of Voter Registration.
8	NC	In some cases, finance committee member decisions are impacted by the theories of modern economics.
	PP	In some cases, the decisions of members of the finance committee are impacted by the theories of modern economics.
9	NC	In current times, commercial agriculture production cost is altered by the location of available land.
	PP	In current times, the cost of production for commercial agriculture is altered by the location of available land.
10	NC	In this study, the corporate tax analysis strategy is pursued by the use of multiple methods.
	PP	In this study, the strategy of analysis for corporate taxes is pursued by the use of multiple methods.

(*Continued*)

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

Table A1. (*Continued*)

11	NC	In many countries, the industrial emission rate balance is controlled by the Agency of Environmental Protection.
	PP	In many countries, the balance of rates in industrial emissions is controlled by the Agency of Environmental Protection.
12	NC	In many countries, consumer protection commerce regulations are enforced by the Commission of Federal Trade.
	PP	In many countries, the regulations of commerce for consumer protection are enforced by the Commission of Federal Trade.

Table A2. List of all items with length 6

1	NC	In present times, the Mexican city skilled labor centralization process is driven by the desire for greater efficiency.
	PP	In present times, the process of centralization of skilled labor in Mexican cities is driven by the desire for greater efficiency.
2	NC	In current times, United States factory employee insurance costs are decreased by the changes in union policies.
	PP	In current times, the cost of insurance for factory employees in the United States is decreased by the changes in union policies.
3	NC	In many countries, modern era general election campaign corruption is created by the hostility of the political atmosphere.
	PP	In many countries, the corruption of campaigns in general elections of the modern era is created by the hostility of the political atmosphere.
4	NC	For our investigation, college graduate career earnings lifetime potential is calculated by the use of statistical analysis.
	PP	For our investigation, the potential of a lifetime of career earnings for college graduates is calculated by the use of statistical analysis.
5	NC	In current times, the online marketplace frozen inventory storage cost is balanced by the strength of popular demand.
	PP	In current times, the cost of storage of frozen inventories for online marketplaces is regulated by the strength of popular demand.
6	NC	In some cases, food product nutrition information labeling laws are drafted by the members of the Codex Committee.
	PP	In some cases, the laws for the labeling of nutrition information on food products are drafted by the members of the Codex Committee.
7	NC	In this study, the married woman financial treatment equality index is developed by the workers of a nonprofit organization.
	PP	In this study, the index of equality for the financial treatment of married women is developed by the workers of a nonprofit organization.
8	NC	In some cases, developing country school choice program significance is judged by the researchers of the new study.
	PP	In some cases, the significance of programs for school choice in developing countries is judged by the researchers of the new study.

(*Continued*)

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

Table A2. (*Continued*)

9	NC	In present times, the European state financial investment return value is increased by the policies of the European Union.
	PP	In present times, the value of returns on financial investments in European states is increased by the policies of the European Union.
10	NC	In this study, national trade business model prediction improvement is anticipated by the introduction of advanced technologies.
	PP	In this study, improvement in the predictions of business models for national trade is anticipated by the introduction of advanced technologies.
11	NC	In current times, high school business marketing education programs are approved by the committee of the relevant organization.
	PP	In current times, programs for education in business marketing in high schools are selected by the committee of the relevant organization.
12	NC	In some cases, discount supermarket soft drink advertising proposals are delivered by an associate of the marketing team.
	PP	In some cases, proposals for the advertising of soft drinks in discount supermarkets are delivered by an associate of the marketing team.

Appendix B Experiment 2 items

Table B1. List of all items with length 4. The parenthesis indicates the parts that need to be removed in order to build the items with length 3

1	NC	In present times, the (currency) inflation constraint action is implemented by the board of the National Bank
	PP	In present times, the action for the constraint of inflation (of the currency) is implemented by the board of the National Bank
2	NC	In many countries, the (factory) automation legislation advice is given by a committee of elected representatives
	PP	In many countries, the advice for the legislation on automation (of factories) is given by a committee of elected representatives
3	NC	In some cases, the (employee) insurance payment reduction is approved by the members of the labor union
	PP	In some cases, the reduction of the payment of the insurance (of the employee) is approved by the members of the labor union
4	NC	In many countries, the (nutrition) fact disclosure regulation is enforced by the Institute for National Health
	PP	In many countries, the regulation of the disclosure of facts (of nutrition) is enforced by the Institute for National Health
5	NC	In other words, the (office) technology adjustment period is anticipated by the introduction of advanced computers
	PP	In other words, the period of adjustment to technology (in the office) is anticipated by the introduction of advanced computers

(*Continued*)

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

Table B1. (*Continued*)

6	NC	In present times, the (alcohol) producer advertisement group is motivated by the competition in Western countries
	PP	In present times, the group for the advertisement of producers (of alcohol) is motivated by the competition in Western countries
7	NC	In certain instances, the (border) taxation consequence summary is approved by the office of foreign trade
	PP	In certain instances, the summary of the consequence of the taxation (at the border) is approved by the office of foreign trade
8	NC	In many countries, the (transit) energy performance standard is controlled by the Agency of Environmental Protection
	PP	In many countries, the standard for the performance of energy (of transit) is controlled by the Agency of Environmental Protection
9	NC	In most situations, the (product) quality assurance department is motivated by the perception of brand performance
	PP	In most situations, the department of the assurance of the quality (of the product) is motivated by the perception of brand performance
10	NC	In some cases, the (highway) construction equipment subsidy is decreased by the changes in infrastructure policies
	PP	In some cases, the subsidy for the equipment for the construction (of the highway) is decreased by the changes in infrastructure policies
11	NC	In other words, the (utilities) monopoly operation condition is driven by the desire for greater efficiency
	PP	In other words, the condition for the operation of the monopoly (of utilities) is driven by the desire for greater efficiency
12	NC	In other words, the (internet) commerce growth expectation is increased by a reform to national patents
	PP	In other words, the expectation of the growth of commerce (on the internet) is increased by a reform to national patents
13	NC	In current times, the (company) acquisition oversight council is regulated by the Committee of Fair Competition
	PP	In current times, the council for the oversight of the acquisition (of the company) is regulated by the Committee of Fair Competition
14	NC	In this study, the (capital) allocation efficiency analysis is pursued by the use of multiple methods
	PP	In this study, the analysis of the efficiency of the allocation (of the capital) is pursued by the use of multiple methods
15	NC	In certain instances, the (welfare) fraud investigation committee is examined by the analysts of financial institutions
	PP	In certain instances, the committee for the investigation of the fraud (of welfare) is examined by the analysts of financial institutions
16	NC	In other words, the (aluminum) shipment application paperwork is increased by the policies of the European Union
	PP	In other words, the paperwork for the application for the shipment (of the aluminum) is increased by the policies of the European Union

(*Continued*)

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

Table B1. (*Continued*)

17	NC	In most situations, the (military) aircraft transaction proposal is delivered by an associate of the marketing team
	PP	In most situations, the proposal for the transaction of aircraft (of the military) is delivered by an associate of the marketing team
18	NC	In current times, the (gender) employment equality movement is supported by the workers of nonprofit organizations
	PP	In current times, the movement for the equality of employment (of genders) is supported by the workers of nonprofit organizations
19	NC	In some cases, the (investment) portfolio diversity strategy is calculated by the use of statistical analysis
	PP	In some cases, the strategy of diversity of the portfolio (of investment) is calculated by the use of statistical analysis
20	NC	In this study, the (business) revenue maximization principle is impacted by the theories of modern economics
	PP	In this study, the principle of the maximization of revenue (of the business) is impacted by the theories of modern economics
21	NC	In certain instances, the (candidate) campaign speech presentation is assessed by the commentators of the news network
	PP	In certain instances, the presentation of the speech of the campaign (of the candidate) is assessed by the commentators of the news network
22	NC	In this study, the (potato) shortage problem management is affected by the location of available land
	PP	In this study, the management of the problem of the shortage (of potatoes) is affected by the location of available land
23	NC	In certain instances, the (healthcare) policy disapproval response is created by the hostility of the political atmosphere
	PP	In certain instances, the response of the disapproval of the policy (for healthcare) is created by the hostility of the political atmosphere
24	NC	In most situations, the (history) education modernization agenda is furthered by the governments of wealthy countries
	PP	In most situations, the agenda for the modernization of education (of history) is furthered by the governments of wealthy countries
25	NC	In many countries, the (machinery) rental agreement negotiation is balanced by the strength of the increasing demand
	PP	In many countries, the negotiation of the agreement of the rental (of machinery) is balanced by the strength of the increasing demand
26	NC	In some cases, the (minority) voter participation statistic is evaluated by the Office of Voter Registration
	PP	In some cases, the statistic of the participation of voters (of the minority) is evaluated by the Office of Voter Registration,
27	NC	In most situations, the (hurricane) relief organization donation is appreciated by the people of impacted communities
	PP	In most situations, the donation to the organization for the relief (of the hurricane) is appreciated by the people of impacted communities

(*Continued*)

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

Table B1. (*Continued*)

28	NC	In present times, the (politics) newspaper coverage commentary is criticized by the president of the United States
	PP	In present times, the commentary of the coverage of the newspaper (on politics) is criticized by the president of the United States

Appendix C Experiment 1 models

Table C1. Experiment 1: First pass duration – region of interest 1

	b	SE	*t*	*p*
Intercept	425.294	12.469	34.108	**<.001**
OPT	3.848	11.782	0.327	.744
DST	−29.376	12.286	−2.391	.017
Length	19.064	18.248	1.045	.296
Type	−2.880	17.285	−0.167	.868
Length:Type	34.139	36.509	0.935	.350

Note: Significant ($p < .01$) findings are indicated in bold.

Table C2. Experiment 1: First pass duration – region of interest 2

	b	SE	*t*	*p*
Intercept	388.056	13.632	28.466	**<.001**
OPT	−3.396	12.464	−0.272	.785
DST	−12.757	12.247	−1.042	.298
Length	−0.790	17.097	−0.046	.963
Type	29.340	16.922	1.734	.083
Length:Type	−45.818	34.484	−1.329	.184

Table C3. Experiment 1: total duration – region of interest 1

	b	SE	*t*	*p*
Intercept	785.050	31.650	24.804	**<.001**
OPT	60.150	27.266	2.206	.027
DST	−126.326	28.930	−4.367	**<.001**
Length	46.216	32.450	1.424	.154
Type	7.657	33.466	0.229	.819
Length:Type	45.218	56.499	0.800	.424

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

Table C4. Experiment 1: total duration – region of interest 2

	b	SE	*t*	*p*
Intercept	849.482	31.542	26.932	**<.001**
OPT	63.417	24.740	2.563	.010
DST	−129.954	26.981	−4.816	**<.001**
Length	17.586	44.005	0.400	.689
Type	−63.951	35.620	−1.795	.073
Length:Type	2.803	80.130	0.035	.972

Table C5. Experiment 1: Regressions onto the critical region. Note that the effect of OPT was driven by a single outlying participant

	b	SE	*z*	*p*
Intercept	1.016	0.099	10.269	**<.001**
OPT	0.256	0.098	2.621	**.009**
DST	−0.099	0.096	−1.031	.303
Length	−0.347	0.072	−4.796	**<.001**
Type	−0.362	0.071	−5.091	**<.001**
Length:Type	−0.082	0.119	−0.690	.490

Table C6. Experiment 1: Regressions onto the critical region (normalized by the length in characters of the critical region). Note that the effect of OPT was driven by a single outlying participant

	b	SE	*z*	*p*
Intercept	−2.881	0.100	−28.812	**<.001**
OPT	0.257	0.098	2.627	**.009**
DST	−0.098	0.096	−1.020	.308
Length	−0.003	0.079	−0.039	.969
Type	−0.124	0.071	−1.751	.080
Length:Type	−0.114	0.120	−0.948	.343

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

Appendix D Experiment 2 models

Table D1. Experiment 2: First pass duration – region of interest 1. Note that the effect of visual WM was driven by a single outlying participant

	b	SE	*t*	*p*
Intercept	426.850	12.490	34.178	**<.001**
OPT	22.320	13.890	1.607	.108
Spelling	−18.750	14.340	−1.307	.191
Length	18.460	15.580	1.185	.236
Type	−12.950	16.850	−0.769	.442
VerbalWM	−21.340	12.160	−1.756	.079
VisualWM	45.800	12.940	3.540	**<.001**
Length:Type	−22.850	23.780	−0.961	.337

Note: Significant ($p < .01$) findings are indicated in bold.

Table D2. Experiment 2: First pass duration – region of interest 2. Note that the effect of visual WM was driven by a single outlying participant

	b	SE	*t*	*p*
Intercept	390.145	13.438	29.034	**<.001**
OPT	6.643	13.883	0.478	.632
Spelling	−36.080	14.897	−2.422	.015
Length	18.161	18.128	1.002	.316
Type	44.968	15.120	2.974	**.003**
VerbalWM	−20.084	12.860	−1.562	.118
VisualWM	33.914	12.637	2.684	**.007**
Length:Type	−9.494	33.492	−0.283	.777

Table D3. Experiment 2: Total duration – region of interest 1

	b	SE	*t*	*p*
Intercept	785.350	27.770	28.285	**<.001**
OPT	38.460	31.450	1.223	.221
Spelling	−22.540	32.360	−0.696	.486
Length	17.510	31.020	0.564	.573
Type	−18.940	33.250	−0.570	.569
VerbalWM	−57.810	28.700	−2.014	.044
VisualWM	48.330	27.750	1.742	.082
Length:Type	−15.400	75.120	−0.205	.838

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

Table D4. Experiment 2: Total duration – region of interest 2

	b	SE	*t*	*p*
Intercept	872.603	33.493	26.053	**<.001**
OPT	84.636	34.828	2.430	.015
Spelling	−92.529	35.970	−2.572	.010
Length	5.502	37.501	0.147	.883
Type	46.029	36.500	1.261	.207
VerbalWM	−39.468	30.778	−1.282	.200
VisualWM	1.813	29.680	0.061	.951
Length:Type	−93.602	74.110	−1.263	.207

Table D5. Experiment 2: Regressions onto the critical region

	b	SE	*z*	*p*
Intercept	1.055	0.108	9.753	**<.001**
OPT	0.050	0.106	0.475	.635
Spelling	0.092	0.113	0.821	.411
Length	−0.272	0.064	−4.276	**<.001**
Type	−0.742	0.076	−9.821	**<.001**
VerbalWM	−0.029	0.097	−0.293	.770
VisualWM	−0.024	0.092	−0.265	.791
Length:Type	0.027	0.143	0.187	.851

Table D6. Experiment 2: Regressions onto the critical region (normalized by the length in characters of the critical region)

	b	SE	*z*	*p*
Intercept	−2.712	0.108	−25.106	**<.001**
OPT	0.056	0.105	0.535	.593
Spelling	0.086	0.112	0.775	.438
Length	−0.011	0.065	−0.174	.862
Type	−0.397	0.076	−5.241	**<.001**
VerbalWM	−0.034	0.096	−0.352	.725
VisualWM	−0.016	0.091	−0.180	.857
Length:Type	−0.003	0.142	−0.024	.981

https://doi.org/10.1017/S0142716424000092 Published online by Cambridge University Press

Appendix E Calculating item information

Calculating information depends on estimating the probability of a word given the surrounding context. A number of methods exist for such an estimation, which have been shown in the literature to correlate with other behavioral measures such as self-paced reading times and ERP patterns, including *n*-gram models (see, e.g., Lopopolo et al., 2017; Willems et al., 2016) and models based on neural networks (e.g., S. L. Frank et al., 2019; Merkx & Frank, 2021). We use Python 3.8.10 and the HuggingFace transformer library (Wolf et al., 2020) version 4.8.2 with OpenAI's GPT-2 neural network-based language model (Radford et al., 2019) accessible through the transformer's interface to make our estimations. Given a sequence of words (e.g., "the cat sat on the") composing the context c, the model produces a probability distribution $P(t|c)$ over all tokens t in its vocabulary. For example, in this case, the probability assigned by the model to the word "mat" is $P(\text{“}mat\text{”}\ |\text{“}the\ cat\ sat\ on\ the\text{”}) = 0.0000000145$ (a very low probability), and the highest probability is assigned to the word "floor" ($P(\text{“}floor\text{”}\ |\text{“}the\ cat\ sat\ on\ the\text{”}) = 0.06707$). To calculate the information in "mat," we use

$$\text{info}(''mat''|''the\ cat\ sat\ on\ the'') \ = \ -\log(P(''mat''|''the\ cat\ sat\ on\ the'')).$$

This formula is known as Shannon's amount of information and is identical to a value known as surprisal in the psycholinguistic literature (cf. Hale, 2001; Levy, 2008).

In order to calculate the information contained in a text segment, we start by tokenizing the text into a sequence of tokens. The library provides its own tokenizer class (AutoTokenizer) which, given a sequence of characters, say, *hello world*, breaks the sequence into tokens that are recognized by the language model. In this case, it would produce the tokens *"hello"* and *"world."*

Out of the original text segment, the tokenization procedure produces a sequence of tokens $t_1, \ldots, t_n$. In order to calculate the amount of information in this sequence, we take advantage of the assumption that information is additive and sum the amounts produced for each token composing the sequence. In particular, we calculate the sum

$$\text{info}(t_1, \ldots, t_n) \ = \ -\log(P(t_1|t_0)) - \log(P(t_2|t_0, t_1)) \cdots - \log(P(t_n|t_0, t_1, \ldots t_{n-1}))$$

where t_0 is a placeholder context used for the first token. In our calculations, t_0 is always the token "the."

For a concrete example, the segment *"the United States factory employee insurance costs"* was tokenized as the tokens "the," " United," " States," " factory," " employee," " insurance," and " costs." Then, we calculated:

$$\begin{aligned}\text{info}(''United\ States\ factory\ employee\ insurance\ costs''\frown) = &-\log(\text{P}(''\text{United}''\frown|''\text{the}''\frown))\\ &-\log(P(''States''\frown|''\text{the United}''\frown))\\ &-\log(P(''factory''\frown|''\text{the United States}''\frown))\\ &-\log(P(''employee''\frown|''\text{the United States factory}''\frown))\\ &-\log(P(''insurance''\frown|''\text{the United States factory employee}''\frown))\\ &-\log(P(''costs''\frown|''\text{the United States factory employee insurance}''\frown)).\end{aligned}$$

Cite this article: Gamboa, J. C. B., Fernandez, L. B., & Allen, S. E. M. (2024). Investigating the Uniform Information Density hypothesis with complex nominal compounds. *Applied Psycholinguistics* **45**, 322–367. https://doi.org/10.1017/S0142716424000092

Investigating the Uniform Information Density Hypothesis in L2 with complex nominal compounds

Anonymous authors

Anonymous affiliation

Abstract

While the Uniform Information Density (UID) hypothesis has been widely investigated for production, little attention has been given to comprehension. In this paper, we extend the study of Gamboa, Fernandez, and Allen (2024), focusing on the reading comprehension of complex nominal compounds (CNC). CNCs are an informationally dense structure that occurs frequently in scientific texts. We follow Gamboa et al., (2024) and compare them with another, less dense structure formed by a noun followed by prepositional phrases (NPP). Based on the UID hypothesis, they predicted CNCs to cause more reading difficulty than NPPs because NPPs spread the information through more words. While Gamboa et al. did find evidence in favor of the UID hypothesis, their pattern of results was unclear. They suggested that this was because reading is less cognitive demanding than production, and that participants were experienced with CNCs. In this paper, recruiting L2 speakers in an attempt to increase the cognitive demand of the reading task. We also investigate experience by comparing Portuguese/Spanish (inexperienced with CNCs) and German speakers (experienced). Our results showed no cross-linguistic difference, and an unclear result pattern, suggesting cognitive demand and experience were not why Gamboa et al. found unclear results.

Introduction

Recently, a substantial number of psycholinguistic research has focused on whether speakers consciously or unconsciously consider the amount of information contained in their utterances (for a review, see Gibson et al., 2019). Research investigating the role of information in language use has suggested that speakers prefer utterances that transfer information at a constant rate. In other words, they prefer to keep the *information density* of their utterances *uniform*, avoiding, for example, producing bursts of too much information at once. This has been shown extensively for production, but there are not many studies investigating whether this is also true for comprehension, and the few studies we are aware of have focused on L1 processing (Collins, 2014; Sikos, Greenberg, Drenhaus, & Crocker, 2017; Meister et al., 2021; Gamboa, Fernandez, & Allen, 2024). In one reading study (Gamboa et al., 2024), the unclear findings led the authors to suggest that reading may not be a hard enough task to evoke the predicted effects. In the present paper, we repeat their study with a population for which reading in English is typically considered to require more cognitive effort, namely, English L2 speakers. The experiments we report take advantage of a particularly dense structure referred to as *complex nominal compound* (CNC) to investigate whether L2 processing is influenced by the amount of information transferred per unit of signal. In addition, by comparing two contrasting groups of speakers that differ in the use of CNCs in their L1, we investigate how these differences may lead their mental models to assign different probabilities to the CNCs in their L2. In the rest of this introduction, we describe the concept of information, define the Uniform Information Density hypothesis, review some of the previous research investigating its predictions on L1 speakers, motivate our predictions for L2 speaker processing of CNCs, briefly discuss the literature on L2 CNC processing, and introduce the two eye-tracking experiments reported in this study.

Information and the Uniform Information Density hypothesis

In the sense used above, "information" is a concept imported from a mathematical theory of data communication (known as Information Theory; Shannon, 1948) where it

is defined in terms of probabilities. For Information Theory, the information content of a given event is operationalized as a number, associated with the event, which depends on its probability of ocurrence. The exact value of this number is related to the notion of *surprisal*: unlikely (i.e., "surprising") events contain a lot of information, and therefore are associated with a big number; likely events are not informative (intuitively, the information about their ocurrence is already available), and therefore are associated with a small number. Crucially, this operationalization of the "amount of information" contained in an event may also be conditioned on other events. For example, after the words "former US President Barack", the probability of finding the word "Obama" is quite high. Therefore, the amount of information associated with "Obama" is lower than if the word "Obama" stood alone. Thus, if an unlikely event becomes very likely because of its context, then its information content (its surprisal) is reduced; conversely, if a likely event becomes very unlikely given the conditioned-on events, then the information associated with it will change to a higher number.

When the theory was proposed, events were originally messages being transmitted by a given device (e.g., a telegraph) through a channel (e.g., a network cable, or radio signals) to a receiver device. To transmit these messages, the transmitter would send a number of symbols through the channel, which, when composed together, would form the message. A well known example of this is Morse code. Each message corresponds to a letter, and is encoded into a sequence of dots and dashes, separated by spaces. Thus, the letter *a* is encoded as • –, the letter *b* is encoded as – • • •, and so on. The spaces, dots and dashes of the Morse code could be (for example) rerepresented as the symbols 0, 1 and 2 respectively. In this way, the letter *a* would be represented by the sequence of symbols 102.[1]

In their applications, both the transmitter and the receiver had knowledge about the probabilities associated with these messages, and by using this knowledge one could define a code (a mapping between symbol sequences and messages) to transmit messages between them in an efficient way. *Efficient* in this sense meant "with the least number of symbols". Following our Morse code example, let us assume that the

messages were English letters. In this case, it is worth noting that not all English letters occur with the same probability in English sentences, i.e., some letters are more common than others. In particular, the letter *e* is very common, while *z* is quite uncommon (see Solso & King, 1976 for a list of the letter probabilities in English). For this reason, it would be smart to associate with *e* a very short code (in Morse code, "1"), while *z* could have a longer code, since it is so uncommon.

The goal of the field, however, was to devise ways to encode the messages not only efficiently, but also in such a way that the transmitted messages could be reliably decoded by the receiver even if some symbols came through the channel incorrectly (e.g., because the network cable was damaged). These errors were generally caused by noise present in the channel, which corrupted the transmitted symbols (e.g., transforming them into other symbols, or deleting some of them). Note that, because each message was associated with a probability, it also carried some amount of information. In our English example, *e* has a high probability of occurrence, and therefore carries less information than *z*, which has a low probability.[2] Shannon showed that the transmitter and the receiver could communicate through the channel with an arbitrarily low amount of errors, as long as the communication did not exceed the channel *capacity*. The channel capacity was the maximum amount of information that could be sent through the channel without errors, and was only dependent on the types of symbols being sent and the amount of noise present in the channel. If the transmitter transferred information above this capacity, there would be no way to guarantee the absence of errors. Thus, efficient and reliable communication was generally achieved by keeping the amount of information sent through the channel at the capacity level at all times (i.e., a constant rate of information transmition).

In Psycholinguistics, early work described a principle of entropy rate constancy (Genzel & Charniak, 2002), later extended as the Uniform Information Density (UID) hypothesis (Jaeger, 2010). According to this principle, communication between speakers could be viewed as a channel, through which a maximum amount of information (its capacity) can be passed per unit of signal (say, time, or number of words), much in the

same way as the channel from Information Theory. Speakers, then, would produce their utterances in such a way that they would be the closest possible to this maximum most of the time, therefore maintaining the amount of information transmitted generally close to a constant. This way, whenever two syntactic structures are available to express the same meaning, a preference would be given to whichever structure would keep the transmitted amount of information per signal unit (i.e., the density) closer to this contant. For example, the work of Jaeger (2010) showed that speakers omit the complementizer *that* more often when it is predictable than when it is not (e.g., *my boss thinks (that) I'm absolutely crazy*), therefore avoiding a trough of information when a complement clause is expected, and avoiding a peak of information when it is not; and Ferreira and Dell (2000) showed a similar effect with *that*-deletion in relative clauses (e.g., *this is the friend (that) I told you about*).

Note that this approach seems to assume that the information present in a word/structure is known by all speakers. It is as if each speaker had access to a central authority that determines the probability of each symbol, and therefore its information content. Of course, this is not how it really works. Different speakers may vary in their mental model, but this variation is typically presumed to be small and therefore is usually neglected in experiments investigating the UID hypothesis and its relation to L1 processing.

A particularly dense structure: Complex Nominal Compounds (CNC)

A nominal compound is a sequence of words that, together, define one single concept (e.g., *university faculty member*, or *stomach ulcer medication*). The last word (the *head*) is always a noun, and the preceding words (the *modifiers*) can be either adjectives or nouns. In this paper, we focus on long NCs containing at least 3 words (which we refer to as *complex* nominal compounds, CNC), and contrast them with another structure composed of a noun followed by prepositional phrases (NPP; e.g., *member of faculty of the university*, or *medication for ulcers in the stomach*), that expresses roughly the same meaning through the use of prepositions.

An important difference between the two structures is how much they spread the information they convey through the signal. If each word is counted as a unit of information,[3] then it is easy to see how an NPP is composed of more information units than CNCs: the same information is conveyed in three words by *university faculty member* and in six words by *member of faculty of the university*. If the final amount of information transmitted by both structures is roughly the same, then we can say that NPPs are less "dense" than CNCs. Indeed, in the experiments below we use items from Gamboa et al., who used a language model to estimate the information of their items, showing that their CNCs do convey on average more information than their NPP counterparts.

Uniform Information Density predictions for Comprehension

If speaker productions conform to the predictions made by the UID hypothesis, it would be natural to expect that the language system would have evolved to also *decode* messages in a similar fashion. However, most of the research related to this hypothesis has focused on speakers' production, and only a few previous studies have focused on comprehension (Collins, 2014; Meister et al., 2021; Sikos et al., 2017; Gamboa et al., 2024), of which two are especially relevant for this paper.

First, Sikos et al. (2017) reports one experiment in which participants read sentences in a G-maze task (Forster, Guerrera, & Elliot, 2009). Sentences began with a frame that was either predictive of the verb object (*The journalist published the...*) or not (*The man evaluated the...*) and contained verb objects that were either pre-modified nouns (e.g., *... carefully written essay*) or post-modified nouns (*... essay that was carefully written*). Sikos et al. then analysed the response times at the object noun (*essay*), predicting that participants would experience less difficulty in predictive than in unpredictive contexts, since the head noun in the predictive contexts was assumed to convey less information. Similarly, they expected less difficulty when the object noun was premodified than when it was post-modified, because the head noun in a premodified context is assumed to convey less information. Their results indicated that

participants indeed processed the head noun (a) faster in predictive contexts than in non-predictive ones; and (b) faster in the pre-modification condition than in the post-modification condition. They argue that both findings constitute indirect evidence for the UID hypothesis from a comprehension point of view. That is, for their first finding, predictive contexts increase the probability of the target word, making it less informative, allowing readers to proceed faster. Intuitively, if the target word causes the information transmission rate to be below the channel capacity, then the reader should compensate by proceeding to the next word. For their second finding, pre-modifiers change the probability distribution of the upcoming head noun, making plausible continuations less informative. In their experiment, the target word was always a plausible head noun.

Second, Gamboa et al. (2024) investigated the UID hypothesis focusing on CNCs. Their two eye-tracking studies compared the reading patterns of English L1 speakers reading English sentences containing either a CNC or an NPP. Their predictions were that participants would find CNCs harder to process than NPPs, arguing that CNCs would transmit a larger amount of information at once. This is justified especially because the structures were presented close to the beginning of the sentence, with little context to support their understanding. Gamboa et al. also manipulated the length of the CNCs in order to assess the impact of peaks of different intensity, predicting that longer items would lead to more difficulty. Examples (1) and (2) illustrate the four conditions of the items of their first experiment: two containing items composed of four words, and two containing items of six words.[4]

(1) a. In present times, **health insurance economy effects** are researched by the analysts of financial institutions (CNC)

b. In present times, **effects of the economy on health insurance** are researched by the analysts of financial institutions (NPP)

(2) a. In some cases, **food product nutrition information labeling laws** are drafted by the members of the Codex Committee (CNC)

b. In some cases, **the laws for the labeling of nutrition information on food products** are drafted by the members of the Codex Committee (NPP)

To measure difficulty, Gamboa et al. counted the number of regressions made onto the critical region (i.e., onto the CNC/NPP), and, since the two structures have considerably different lengths, measured the time participants spent looking at the text segments *following* the critical region. The assumption was that longer times and more regressions meant more reading difficulty. In particular, after the critical region, the target sentences contained two segments that were separately analysed: a passive construction (*are researched*, *are drafted*), and its agent (*by the analyst*, *by the members*).

In both experiments, they found more regressions toward the critical region after longer structures. In their second experiment, they also found longer reading times for the segment denoting the agent. Both findings were interpreted as evidence in favor of the UID hypothesis, under the assumption that longer structures, and CNCs represent peaks in the information density of the communication signal.

However, they did not find significant effects in the majority of the analysed reading measures. They suggested two reasons why this was the case. First, they noted the role of experience in the processing of CNCs: their participants were mostly university students who may be more used to reading this type of structure than the general population. A population with less experience with CNCs may have behaved differently. Second, they pointed out that reading is much less resource-intensive than either production or the G-Maze task used by Sikos et al. (2017). They suggested that increasing the task's associated cognitive load could lead to a clearer pattern of results.

In order to explore these two suggestions, in this study, we repeat their experiments with speakers for which these arguments may not hold, namely, L2 speakers of English. While we do recruit participants from a university environment, we still presume that reading in a second language imposes higher cognitive load demands than reading in a first language.

UID in the Second Language

As mentioned earlier in the discussion on Information Theory, both the transmitter and the receiver need to have knowledge of the probabilities associated with each message, and most UID experiments (focusing on L1) just assume that word probabilities are "agreed" between, or independent of, speakers. But if these probabilities are high in the mental model of a certain speaker and low in the mental model of another speaker, some problems might occur: the former might believe they are sending a constant amount of information while the latter may perceive certain portions of the signal as more or less dense, potentially finding them hard to process. While L1 speakers may indeed have roughly the same probabilistic model of word occurrences, this is most likely not true for non-native speakers whenever their second language differs substantially from their first language in the use of either a set of words or a given structure. Of course, the resulting difficulty might be modulated by the amount of resources available for each individual. For example, it is possible that individuals with a higher working memory capacity would deal with information peaks more easily than those with a lower capacity. Similarly, proficiency or vocabulary in the second language is likely to have an influence on how hard certain structures are perceived to be, and on the probability speakers assign to each word or structure.

In this paper, we explore the impact of the aforementioned 'mental model' by recruiting L2 speakers with different L1s that make different use of CNCs. We recruit two groups of English L2 speakers: one composed of L1 speakers of languages that rarely use (and therefore are less experienced with) CNCs – Portuguese and Spanish (P/S) –, and one composed of L1 speakers of a language that does use them often – German. In addition, we take into account the aforementioned individual differences, namely, working memory and proficiency.

L2 processing of CNCs

Most of the existing work in (C)NC processing in L2 has not taken into account this information theory perspective. If L1 speakers experience difficulty with the peaks of

information caused by CNCs, it is natural to expect that L2 speakers will experience even more difficulty, due to their more limited processing ability in the L2 compared to that of native speakers. Additionally, it is widely assumed that L2 speakers at all but the most advanced levels have more difficulty in linguistic processing than their native speaker counterparts, although accounts differ for why this is the case (Clahsen & Felser, 2006; Hopp, 2010; Jiang, 2007; Kaan, 2014; McDonald, 2006). In general, due to their comparative lack of experience with the language, L2 speakers dedicate more processing resources than native speakers to the basic task of parsing sentences, and/or parse sentences at a more surface level than native speakers, and thus are less able to integrate detailed information into their parsing during real-time processing.

Although several studies have investigated L2 speakers' comprehension of CNCs in behavioral studies using paraphrasing, definitions, and translations (e.g., Bartolic, 1978; Carrió Pastor, 2008; Carrió Pastor & Candel Mora, 2013; Horsella & Pérez, 1991; Olshtain, 1981; Pabón Berbesí & Domínguez, 2009; Trimble & Trimble, 1977; van Helmond & van Vugt, 1985), none have investigated it using on-line methods that would provide more information about real-time processing difficulties, and none have explicitly compared groups of L2 speakers with contrasting L1s.

Present Study[5]

In this paper, we assess the processing of English L2 speakers using the materials and methodology of Gamboa et al. (2024), comparing two sets of speakers with L1s that differ particularly on their use of CNCs. As mentioned earlier, our rationale is that L2 speakers who speak L1s where CNCs are rare (e.g., Spanish, Portuguese: Carrió Pastor, 2008; Jullian, 2001; Oliveira et al., 2006; Richman, 1969) will perceive CNCs as more informationally dense than will L2 speakers with L1s where CNCs are pervasive (e.g., German: Berg, 2012, 2016), because the latter group have more experience overall with strategies for comprehending dense structures of this type, and because their mental models assigns a higher probability to this kind of structure, making it comparatively less dense.

We report findings of two eye-tracking experiments, both of which are very similar: participants read sentences containing either a CNC or an NPP, which were additionally manipulated in length. In Experiment 1, CNCs had either 4 or 6 words, but were not fully factorial, limiting direct comparisons between them; in Experiment 2, we overcome this limitation by using items with 3 and 4 content words that were fully factorial.

For both experiments, our hypotheses are similar to those of Gamboa et al. (2024). We expect participants to have more difficulty with CNCs than with NPPs, reflecting the higher information density of the CNCs. We also expect participants to have more difficulty with longer structures than shorter structures, since longer structures will have a higher information density overall. In particular, for both experiments, participants will regress more often, as well as produce longer reading times in sentences containing CNCs than in sentences containing NPPs. In addition, this behavior will be modulated by their previous experience with CNCs in their native language. That is, we expect P/S speakers to have more difficulty when processing CNCs than German speakers, yielding an interaction between speaker L1 and the type of structure (CNC/NPP) present in the sentence. Finally, as discussed above, we expect the aforementioned effects to be affected by participants' working memory abilities and proficiency levels, so that higher working memory and proficiency scores would lead to shorter reading times overall.

EXPERIMENT 1

Method

Participants

Participants were recruited from the community of the [**removed for review**]. All participants had normal or corrected-to-normal sight and were compensated with payment or course credit.

In the German-speaking group, we recruited 35 German native speakers (15 female, 20 male; mean age: 25 years; sd: 3.5) all of which were from Germany. In the Portuguese/Spanish speaking group, participants were 33 native speakers (15 female, 18

male; mean age: 27 years; sd: 4.1) of either Portuguese (10 participants, from Brazil and Portugal) or Spanish (23 participants, from Chile, Colombia, Ecuador, Honduras, Mexico, Spain and Venezuela).

Materials

The materials were the same as those of Gamboa et al.'s (2024) Experiment 1.

Eye-tracking reading task. Participants read sentences and answered easy comprehension questions while having their eye movements recorded by an eye-tracker. Critical sentences (see Table 1) contained either a CNC or an NPP composed of either 4 or 6 content words. Each critical sentence was divided into five parts: a *preamble* containing three words and ending with a comma, a *critical region* containing the critical structure, a post-critical *region of interest 1 (ROI1)* composed of a passive construction, a spillover *region of interest 2 (ROI2)* containing its agent, and a *wrap up* region ending the sentence. Given CNCs N1-N2-N3-N4 and N1-N2-N3-N4-N5-N6, corresponding to lengths 4 and 6, respectively, the associated NPPs had structure N4-P-N3-P-N1-N2 and N6-P-N5-P-N3-N4-P-N1-N2, respectively. Given the sentence length, item presentation was such that the preamble and the critical region were on the first text line, and the remaining regions were on the second text line. See Gamboa et al. (2024) for detailed information about item creation and norming.

Oxford Placement Test Part 1 (OPT). As a measure of English proficiency, participants answered the OPT, composed of 50 sentence frames containing a gap, along with three suggested completions, only one of which results in a grammatical sentence. In each case, the participant's goal is to pick the correct alternative.

Language Background Questionnaire (LBQ). Participants additionally filled out a language background questionnaire containing questions about their language use, their self-assessed proficiency, and the contexts in which they used each language. The collected data was used to ensure that participants were not significantly exposed to English before the age of 5.

Digit span test (DST). Participants' working memory was assessed with a DST. Participants saw sequences of digits and had to recall them in their correct order.

Sequences progressed from 2 to 9 digits. A participant's final score was the longest length for which they recalled two sequences correctly.

Design

Items of length 4 were not related to items of length 6. Thus, two counterbalanced presentation lists were constructed so that each list contained six NPPs and six CNCs of each length.

Each participant read four practice sentences, followed by a set composed of 24 target sentences and 40 fillers, with a break halfway through the set, producing two blocks of 32 trials, presented in a random order per participant. Practice and filler sentences were similar to target sentences, but contained no long CNC or NPP. Each sentence was followed by a comprehension question (see bottom of Table 1), which participants answered by pressing the letters X or M. Questions were easy to answer to prevent participants from developing strategies which could influence their reading patterns.

Apparatus

The Experiment Builder software from SR Research was used to program the eye-tracking task, which was run with a color monitor at a resolution of 1280 x 1024 and an EyeLink 1000 recording the right eye at 1000Hz. Participants read with both eyes at approximately 100cm from the monitor, using a chin rest in order to keep their head stable. Stimuli were presented with font Courier New. Eye-tracker calibration was performed before the task, after the break, and as needed throughout the study. Drift correction was used with a fixation cross at the beginning of every trial.

Procedure

After signing the consent form, participants were administered the OPT, followed by the LBQ, the DST and the eye-tracking task. They were informed that the eye-tracking sentences were related to economics, and were always followed by a question. Sentences started with a drift correct at the position of their first letter. Participants were

instructed to read at their own pace, and then press spacebar in order to proceed to the comprehension question. In the eye-tracking task, critical sentences were shown with the preamble and the critical region in the first line of text, and the rest of the sentence in the second line.

Instruction to all computer tasks was given verbally and in written form before the task and after the practice trials.

Analysis

Eye-tracking data preprocessing

Prior to analysis, trials with incorrect answers to the comprehension question were discarded. From the remaining data, a fixation report was exported containing the position and duration of the fixations performed on each trial. Following the same procedure as in Gamboa et al. (2024), we reduced systematic noise by running their fixation position correction algorithm. We additionally removed trials that were too noisy prior to correction (for which we expected the correction procedure would produce results that looked reasonable, but that did not reflect the participant's reading behavior), or that remained too noisy after correction.[6] (see Table 2).

Each word was associated with an interest area around it. Fixations positioned inside a word's interest area were treated as fixations on that word. Using the corrected data, we extracted the same reading measures analysed by Gamboa et al. (2024) in ROI1 and ROI2 for each trial.[7] These were **First pass Duration (FPD)**, the sum of the duration of all fixations made on a region when the reader first arrives at it from the preceding text, **Total Duration (TD)**, the sum of the duration of all fixations made on a region, and the number of **regressions onto the critical region (Reg2CR)**, fixations on a word of the critical region (say, *health* in Table 1) whose previous fixation was positioned inside the interest area of a subsequent word, regardless of whether the previous fixation was inside the critical region (e.g., *economy*) or outside of it (e.g., *researched*).

In order to analyse the extracted measures, and following Gamboa et al. (2024)'s

procedure, we fit four models for the duration measures (FPD and TD), and one additional model for the analysis of Reg2CR. Because we are analysing five reading measures, models were Bonferroni corrected so that only p-values smaller that $0.05/5 = 0.01$ were considered significant (von der Malsburg & Angele, 2017).

Analysis of the duration measures

Before analysis, trials with extreme measure values were trimmed. During the analysis of FPD, trials with FPD shorter than 80ms or longer than 1000ms were discarded; during analysis of TD, trials with TD shorter than 80ms or longer than 2000ms were discarded (see Table 3).

We fit a separate generalized mixed model for each duration measure (FPD in ROI1, FPD in ROI2, TD in ROI1 and TD in ROI2) using a Gamma distribution with the *identity* link function using the R language (R Core Team, 2013) and the lme4 package (Bates, Mächler, Bolker, & Walker, 2015). The calculated p-values are produced using the lmerTest package (Kuznetsova, Brockhoff, Christensen, et al., 2017). Models were maximally specified (Barr, Levy, Scheepers, & Tily, 2013) with random effects for subjects and items and fixed effects including the structure type (CNC vs. NPP), length (4 vs. 6) and language (P/S vs. German), as well as the OPT scores, and the DST scores. Factor variables were sum coded and continuous variables were scaled and centered.[8]

Analysis of the regression count

For the regression count, we follow Gamboa et al. (2024) and fit a maximally specified generalized mixed model passing *family = poisson(link=log)* using the same fixed and random effects as in the other models.

Results

Accuracies were high in both groups and are not further discussed (see Table 2). The German group scored a mean of 76.97% (SD: 11.73) on the OPT and a mean of 6.57

out of 9 (SD: 1.12) on the DST. The P/S group scored a mean of 74.42% (SD: 11.05) on the OPT and a mean of 6.85 (SD: 0.87) on the DST.

Turning to the eye-tracking data, Table 3 shows the number of trials trimmed before analysis of the duration measures, and Table 4 shows the average value for each measure. There were no significant effects for FPD in either ROI1 or ROI2 (Figure 1a and 2a, respectively). For TD, we found a significant effect of structure type in ROI1 ($t=3.828$, $p<.001$), indicating that CNCs took significantly longer to process than NPPs for both groups; and of language ($t=2.579$, $p<0.005$), indicating that German speakers gazed significantly longer at ROI1 than P/S speakers (see Figure 1b). For ROI2, we found no significant effects for TD. As for Reg2CR (Figure 3), P/S speakers regressed significantly more toward the critical region than German speakers ($z=-3.130$, $p<.005$). In addition, NPPs ($z=-5.802$, $p<.001$) and longer structures ($z=-6.138$, $p<.001$) led to more Reg2CR than CNCs and shorter structures, respectively. Full model results are available in Appendix A. We found no effects of proficiency or DST in any reading measure.

Discussion

Experiment 1 followed the design used by Gamboa et al. (2024) in an attempt to investigate the processing of CNCs and NPPs based on the predictions made by the Uniform Information Density (UID) hypothesis. Participants read sentences containing either a CNC or an NPP, and these were composed of either 4 or 6 words. In order to increase the cognitive load associated with the processing of CNCs, we recruited L2 speakers of English. Furthermore, in order to explore the impact of experience in the processing of CNCs, these speakers were L1 speakers of Portuguese and Spanish (P/S) and of German. We predicted that CNCs would be perceived as harder to process than NPPs, and that this difficulty would be starker for the P/S group, given that it is not a common structure in their L1. We also predicted that longer structures would lead to more difficulty, and that the difficulty would be modulated by working memory (WM) abilities and by proficiency scores, as measured by a digit span test (DST) and the

Oxford Placement Test (OPT), respectively.

Our findings do suggest that CNCs were harder to process than NPPs: we analysed duration measures extracted from the regions succeeding the manipulated structure (First Pass Duration, FPD; and Total Duration, TD), and found longer total durations (TD) in the region of interest 1 (ROI1). In addition, participants also regressed more after 6-word items than after 4-word items, in accordance with the more prominent peak in information associated with 6-word items. We interpret these finding as evidence in favor of the UID hypothesis.

However, our findings are not completely consistent with the UID hypothesis. First, none of the duration measures was significantly affected by structure length. We believe this is because of the large variability presented by the measures, which was observed regardless of structure length (Figures 1 and 2), and may also explain why measures were not significantly affected by individual differences such as proficiency or working memory. Second, participants also, perhaps unexpectedly, regressed significantly more in sentences containing NPPs than CNCs. We come back to this latter counterintuitive result in the General Discussion.

Turning to the differences between the two languages, contrary to our expectations, we found no structure:language (nor length:language) interaction. That is, experience with CNCs did not play a particular role in determining reading difficulty. Interestingly, however, while German speakers gazed significantly longer at ROI1 than P/S speakers, P/S speakers performed more regressions than German speakers, possibly reflecting different reading strategies used by speakers of the two language groups.

Our findings, however, are limited in a number of other ways which are discussed by Gamboa et al. (2024) and also apply to the L2 case. First, it is not clear the extent to which the presence of adjectives in the items may have helped with the processing of CNCs. For example, upon seeing the CNC ***modern*** *era* ***general*** *election campaign corruption*, speakers may be led to expect a noun after *modern* and *general*. This may have also disproportionately facilitated the CNC processing for the P/S group: While noun+noun (e.g., *business model*) structures are not common in Portuguese or Spanish,

adjective+noun structures are ubiquitous, allowing for several adjectives to follow one another (e.g., *mercado financeiro/financiero especulativo internacional* → *international financial speculative market*). Second, the experimental design is not ideal to compare the two lengths (4 and 6) used in the experiment, since the items used with each length are completely different. In order to be able to compare lengths, it would be useful to have a fully factorial set of items. Third, many of the content words used in Experiment 1 could be used as verbs, potentially leading participants to be garden pathed. Fourth, the items contained a number of collocations (e.g., *United States*), which may again have helped participants in their processing. Finally, when measuring WM, it was not clear whether the DST measured *verbal* or *visual* WM, and it would be beneficial to find a way to disentangle these two storage units (see, e.g., Baddeley, 2011, for a review on the WM construct). Hence, in the following, we attempt to address the aforementioned limitations by conceptually replicating Gamboa et al.'s Experiment 2.

EXPERIMENT 2

In Experiment 2, items do not contain collocations and are composed solely of nouns that cannot be used as verbs. In order to make lengths comparable, we use items of length 3 and 4: length-4 items are built by adding a single word to length-3 structures, such that any difference in behavioral measures between the two can only be attributed to the additional word. Instead of the DST, we measure WM with a verbal and a visual serial order reconstruction task (Jones, Farrand, Stuart, & Morris, 1995, SORT), reflecting the phonological loop (used for processing language) and the visuo-spatial sketchpad (used for visual processing and typically assumed not to be involved with the processing of language), respectively (for more details on the WM construct, see Baddeley, 2011). In addition to measuring proficiency with the OPT, we also use a misspelling identification task (MSIT) to measure quality of lexical representation (Perfetti & Hart, 2002), since these have been shown to explain different variance in priming and eye-tracking data (see, e.g., Andrews, Veldre, & Clarke, 2020). Finally, we decided to recruit only Spanish speakers for the P/S group (here renamed as "Spanish")

in an attempt to reduce the noise associated with different L1s.

As before, following the UID hypothesis, we predict that CNCs and longer items would lead to longer duration measures and more regressions than NPPs. We also predict an interaction between L1 and structure type, and that reading measures would be significantly influenced by proficiency and verbal WM scores, but not by visual WM scores.

Method

Participants

In Experiment 2, we do not include Portuguese speakers. Participants were 31 Spanish L1 speakers (12 female, 19 male; mean age: 26 ± 5.6 years old) from Bolivia, Colombia, Costa Rica, Ecuador, Mexico, Peru, Puerto Rico, Spain and Venezuela, and 29 German L1 speakers (9 female, 20 male; mean age: 26 ± 6.1 years old) from Germany, all of which were recruited at [**removed for review**] and received either course credits or 15 Euros for their participation. All participants had normal or corrected-to-normal sight.

Three Spanish and one German speakers were subsequently discarded because they were exposed to another language before the age of 5. Additionally, one German speaker reported to be dyslexic in the Language Background Questionnaire and was removed from analysis; and one German speaker could not be calibrated with the eye-tracker and was also removed from analysis. Below we report the data of the remaining 28 Spanish speakers and 26 German speakers.

Materials

The materials were the same as those used in Experiment 2 of Gamboa et al. (2024). Participants filled out the language background questionnaire, then performed an eye-tracking reading task, followed by OPT, the misspelling identification task, and two serial order reconstruction tasks assessing their visual and verbal working memory.

Eye-tracking reading task. As discussed above, the materials are similar to those of Experiment 1, with a CNC/NPP appearing close to the beginning of each

critical item. Items were created with CNCs that are composed exclusively of nouns, which cannot be used as verbs, and were checked for collocations by three English native speakers.

The items are fully factorial (see Table 5), allowing for a clearer comparison between lengths 3 and 4. Four-word CNCs were created by taking the 3-word CNCs and preceding them with an additional noun (e.g., inflation contraint action → *currency* inflation contraint action). NPPs were created from the CNCs.

Language Background Questionnaire (LBQ). This was the same as used in Experiment 1.

Oxford Placement Test Part 1 (OPT). This was the same as used in Experiment 1.

Misspelling identification task (MSIT). Participants received a list of 215 words and were tasked with circling those that were incorrectly spelled (50 out of the 215). Following Lemhöfer and Broersma (2012), participants' final score was calculated as $\frac{1}{2}\left(\frac{100}{50}W + \frac{100}{165}NW\right)$, where W and NW indicate the number of correctly identified *words* and *nonwords*, respectively.

Serial Order Reconstruction Task (SORT). Participants were additionally administered two SORTs, assessing their visual and verbal working memory. In both tasks, participants saw sequences on the screen (visual: 7 dots in different locations; verbal: 8 letters at the center of the screen). After the end of the sequence, all elements reappeared on the screen, and participants were tasked with selecting the elements in the order in which they had appeared. There were 3 practice and 20 critical trials.

Design

Each item had four versions, and four presentation lists were created forming a Latin square design. Each participant only saw one version of each item.

The eye-tracking reading task was divided into a first block containing 4 practice sentences, plus two randomized blocks, separated by a short break, containing a total of 68 trials (40 fillers and 28 critical trials). Fillers contained no long CNC or NPP. Each

sentence was followed by a comprehension question and the participant answered by pressing the letters X and M (see bottom of Table 5).

Apparatus

We used the Experiment Builder software from SR Research to program the eye-tracking task. The task was run with a color monitor at a resolution of 1280 x 1024, font Courier New, and an EyeLink 1000 recording the right eye at 500Hz. Presentation was binocular at approximately 90cm using a chin rest. Calibration was performed before the task, after the break, and as needed. Drift correction was used with a fixation cross at the beginning of every trial.

Procedure

Participants started by signing the consent form, after which they were administered the OPT and the misspelling identification task in paper.

They then proceeded to a computer where they performed the eye-tracking task, the visual SORT and the verbal SORT. As in Experiment 1, the critical sentences in the eye-tracking task were shown with the preamble and the critical region on the first line, and the rest of the sentence on the second line.

Analysis

Eye-tracking data preprocessing

As in Experiment 1, the gaze data was corrected with the correction algorithm described in Gamboa et al. (2024), and trials that were either too noisy before correction or too noisy after correction were excluded. From the corrected data, we extracted FPD and TD for ROI1 and ROI2, as well as Reg2CR. In total, we fit 5 models, one for each of the extracted measures. To account for multiple comparisons, we consider significant only p-values smaller than $0.05/5 = 0.01$.

Analysis of the duration measures

For the duration measures, we fit separate maximally specified Gamma distributed generalized mixed models with the *identity* link function, using structure type (CNC vs. NPP), length (3 vs. 4) and language (Spanish vs. German) as our main predictors (all of which were sum coded), as well as the OPT score, the misspelling identification task score, and the two working memory scores as covariants (all covariants were scaled and centered).[9]

Analysis of the regression count

Regression counts were analysed with generalized mixed effects models with the Poisson distribution and the log link function (i.e., passing *family = poisson(link=log)*, using an effects structure similar to that of the duration measures.

Results

Due to a problem with item presentation, one item (item 22) was removed from the data. As in Experiment 1, accuracies were high and are not considered further (see Table 6). Spanish L1 speakers scored a mean of 37.4 out of 60 (SD: 5.78) on the OPT, and a mean of 76.6 out of 100 (SD: 7.11) on the MSIT. German L1 speakers scored a mean of 41.4 (SD: 4.98) on the OPT, and a mean of 77.6 (SD: 9.67) on the MSIT. Visual and verbal SORT scores were correlated both for the Spanish group ($r=0.424$, $t=2.389$, $p<.05$) and for the German group ($r=0.544$, $t=3.176$, $p<.005$).

Turning to the eye-tracking data, Table 7 shows the number of trials trimmed before analysis of the duration measures, and Table 8 shows the average value for each measure. Figures 4 and 5 show the duration measures for ROI1 and ROI2, respectively, for each condition and group. We found a significant effect of spelling for FPD (see Figures 7a and 7b) both in ROI1 ($t=-3.077$, $p<0.005$) and in ROI2 ($t=-5.274$, $p<0.001$), such that better spelling led to shorter FPD and TD. In addition, we found a significant effect of verbal WM for TD (see Figures 7c and 7d) both in ROI1 ($t=-3.147$, $p<.005$) and in ROI2 ($t=-3.173$, $p<.005$), such that higher verbal WM scores led to shorter TDs.

Figure 6 shows the Reg2CR for each condition and group. We found an effect of Spelling (t=3.035, p<.005), such that better Spelling led to **more regressions**. As can be seen in Figure 7e, this effect seems to be driven by the German group, with the Spanish group keeping a relatively constant number of regressions regardless of their spelling scores. We found an effect of length (t=-3.209, p<.005) and structure type (t=-9.982, p<.001), such that longer structures and NPPs led to higher regression counts. Full model results can be found in Appendix B. In order to better evaluate the relation between the significant predictor covariates and the reading measures, the graphs from Figure C1 were also produced with cubic splines and are available in Appendix C.

Following Gamboa et al. (2024), we additionally used R's *car* package (Fox & Weisberg, 2019) to calculate Variance Inflation Factors[10] (VIF) for all models to ensure that models were not affected by any potential correlations between the included variables. Despite the significant correlation between visual and verbal WM scores reported above, all VIF scores were lower than 2.

Discussion

Experiment 2 proceeded with our investigation of the UID hypothesis in L2 by addressing a number of limitations of Experiment 1. As in Experiment 1, we predicted CNCs and longer structures to lead to longer reading measures and more frequent regressions, and expected Spanish speakers to have more difficulty in CNC processing than German speakers, as reflected by an interaction between L1 and structure type.

Contrary to the predictions derived from the UID hypothesis, we found no effect of structure type, length or language for any of the duration measures, despite the more controlled items used in this Experiment. As for the regressions, Experiment 2 results mirror those of Experiment 1: participants regressed more toward longer vs. shorter structures (as predicted) and more toward NPPs than for CNC (not as predicted). We discuss why this was the case in the general discussion, where we also compare our findings with those of Gamboa et al. (2024).

When it comes to the language differences, Experiment 2 yielded no structure:language interaction, indicating, again, that experience with CNCs did not substantially affect reading difficulty. In addition, we observed no significant effect of language in the regression model, raising doubts on whether the two groups do use different reading strategies.

Finally, as opposed to Experiment 1, in Experiment 2 FPDs were affected by spelling, and TDs were affected by verbal WM, potentially reflecting the more controlled items we used in this experiment. However, better spellers also regressed more. This is unexpected, and we come back to this result in the General Discussion.

General Discussion

This study aimed to contribute to the body of research investigating the Uniform Information Density (UID) hypothesis from the point of view of comprehension, a perspective that has been given little attention relative to production. We followed the operationalization of Gamboa et al. (2024) and compared the processing of complex nominal compounds (CNC), structures that arguably constitute peaks of information density, with that of nouns followed by prepositional phrases (NPP), which spread the information through more words. We also used two different structure lengths (4 vs. 6 in Experiment 1; 3 vs. 4 in Experiment 2) in order to compare the different degrees of density evinced by these structures, and included covariates measuring English proficiency and working memory (WM). Recall that Gamboa et al. (2024) explained their unclear pattern of results by appealing to the differences between production, a resource-demanding task, and comprehension, which tends to be less resource-intensive. In an attempt to make our eye-tracking reading task require the engagement of more cognitive resources, we recruited English L2 speakers. We further compared the reading behavior of speakers of languages that do not typically use CNCs (Portuguese and Spanish in Experiment 1; Spanish in Experiment 2) with that of speakers of a language that does (German), not only looking for potential cross-linguistic influences emanating from their L1, but also exploring the UID prediction that different L1 speakers may

have different intuitions about what constitutes a peak in information density. Thus, we predicted that participants would regress more from and look longer at the text segments succeeding the critical region when it contained a CNC than an NPP, and that English proficiency and working memory would be significant predictors of this reading behavior. In addition, we expected Portuguese/Spanish (P/S) L1 speakers to be more affected by CNCs than German L1 speakers.

Table 9 shows a summary of our results for the critical region, the region of interest 1 (ROI1) and the region of interest 2 (ROI2), and also provides a comparison with the L1 results reported by Gamboa et al. (2024). As predicted by the UID hypothesis, participants did regress more in both experiments towards longer structures, a finding that is in line with the L1 results. In addition, participants looked longer in ROI1 after CNCs than after NPPs in Experiment 1, but this effect was not replicated in the better controlled Experiment 2.

Apart from these results, the two experiments reported here have in common with Gamboa et al. (2024) the fact that they present very little evidence in favor of the UID hypothesis. That is, despite the new significant effect of structure type on Total Durations (TD) in ROI1 in Experiment 1, we were unable to reproduce the main effect of First Pass Duration (FPD) in ROI2 found in their Experiment 2. Indeed, against our predictions, but replicating their findings, participants also regressed more after NPPs than CNCs. Gamboa et al. (2024) suggest that this might be an artifact of the head noun position in the two structures. They argue that, presumably, participants have to regress after an NPP in order to recall its head noun and parse its agreement with the tense in ROI1, since it is far away from the tense verb. In the CNC condition, they read the head noun right before the ROI1, so this regression is not needed.

L1 speakers of both P/S and German were roughly equally affected by the two structures in L2 English, as evidenced by the lack of an interaction between structure type and language in both experiments. That is, the considerably different use of compounds in the speakers' L1 did not lead to a measurable difference in the processing of compounds in their L2 (English). In addition, the two groups differed in Experiment

1: P/S speakers regressed more often than the German speakers, while German speakers looked longer at ROI1 than P/S speakers. It is unclear why these differences between groups emerged, and we hesitate to interpret too much from these findings: they were not present in Experiment 2, suggesting they may have been spurious results that were mitigated by the better controlled items employed in that experiment.

Still, if we *were* to try interpreting this effect, one possibility is that they reflect cultural differences in reading patterns of the different populations. That is, maybe German readers simply adopt a more careful reading strategy, gazing longer at the regions that cause them difficulty, while P/S speakers adopt a 'riskier' strategy, proceeding more quickly in the hope that the subsequent words will clarify the overall meaning of the sentence, but then regressing more often. Further research is necessary in order to confirm this explanation and shed more light on these findings.

Turning to the individual differences, the effects we found in this study were substantially different from those of Gamboa et al. (2024). In particular, we were unable to replicate the working memory effect they found for TDs in Experiment 1; and only found working memory effects in Experiment 2, the experiment in which they had found no such effect. These 'new' working memory effects were as predicted: while *visual* working memory had no effect in reading measures, participants with better *verbal* working memory spent significantly less time in ROI1 and ROI2.

In addition, in Experiment 2, spelling had a major impact in all regions analyzed. Participants who performed better at the spelling task had shorter FPD in ROI1 and ROI2. Surprisingly, however, they also regressed more. That is, the better the quality of lexical representations possessed by speakers, the more regressions they performed towards the critical region (regardless of whether it contained a CNC or an NPP). It is unclear why this happened. For example, perhaps the better spellers had better intuitions about the word combinations used in English, and thus had difficulties with the items because they did not contain collocations. Or perhaps the high cognitive demand caused by the target structures led them to regress more to recheck the words. Further research is needed to investigate this effect.

One of the main contributions of this study is the way we derived our predictions from the UID hypothesis for the L2 speakers. As discussed earlier, the amount of information associated with a given symbol (word, phoneme, structure, etc.) is dependent on its probability. Without this probability, it would be impossible to know the amount of information of the symbol. In most studies devoted to investigating the UID hypothesis, this probability has been treated as absolute, as if given by a central authority that speakers are capable of accessing whenever they want the probability of a symbol. While this may work for L1 speakers, L2 speakers may differ drastically on the probability they assign to each symbol. That is, a German speaker, used to compounds, may assign them a much higher probability than a Spanish speaker, in whose language compounds are not common. We believe this is a useful conceptualization for studying the UID hypothesis in the L2.

While our results are somewhat different from those of Gamboa et al. (2024), both experiments are consistent in their relatively weak support for the UID hypothesis.[11] In other words, it seems that the lack of findings in their experiments does not stem from the cognitive costs associated with the reading task, or, if it does, that reading in a second language is still not as demanding as we had predicted, and that a still harder task (say, the G-Maze task used by Sikos et al., 2017) may be needed to evince results aligned with the UID hypothesis.

Alternatively, it is also possible that experience with CNCs may have played a stronger role than previously assumed in hiding the predicted effects. Note that the participants in the two experiments we reported here were mostly University students (many were Master's students), who were presumably used to reading scientific papers in English, a register in which compounds are especially common (see Biber & Gray, 2011). In addition, the P/S group in Experiment 1 and the Spanish group in Experiment 2 was composed of participants living in Germany, who had learnt at least *some* German and were at least *partially* used to reading compounds in their everyday lives.[12] Therefore, it might be that recruiting L2 speakers from a different environment, who do not frequently read CNCs, would lead to clearer effects of structure type and

length.

Repeating the experiments with such a less experienced group of P/S (or Spanish) speakers may also be an interesting way to examine how robust such a cross-linguistic effect (from an L1 to an L2) may be. In that case, finding an interaction between language and structure type or between language and length (that we did not find in our experiments) would suggest that it was indeed the participants' experience with these structures that led to our null findings and blended the two groups together. Finding *no difference*, on the other hand, would also be interesting, given how starkly the two groups of L1s differ in their use of compounds.

Finally, Gamboa et al. (2024) discuss a number of technical limitations that might also have affected the results, and that also apply to the L2 case. First, the *a priori* choice of eye-tracking measures was somewhat arbitrary (one early measure and one late measure), and it is not clear whether we would have found a clearer pattern of results had we chosen a different set of measures. Second, because CNCs and NPPs vary substantially in length, they opted for a spill-over based study design in which they analysed the subsequent text segments. While this is not an uncommon choice (e.g. Christianson, Luke, Hussey, & Wochna, 2017; Jared & O'Donnell, 2017; Paape & Vasishth, 2022; Pickering & Traxler, 1998) for the analysis of eye-tracking measures, it may partly explain the aforementioned inconsistencies between studies and experiments – note that Reg2CR results, analysing the critical region itself, were much more consistent. Third, because of the length of the sentences, these subsequent segments were presented in a separate text line. This may have further impacted the duration measures, since participants had to perform a long return saccade from the end of the first text line to the beginning of the second, often fixating in unintended positions. These limitations may have contributed to the weak findings of this study.

Despite these limitations, the results reported here suggest that the Uniform Information Density hypothesis may not hold for comprehension as well as it does for production. Further research is needed to better understand in what situations it does hold, if any.

Conclusion

In this paper, we investigated the predictions of the Uniform Information Density (UID) hypothesis from the point of view of comprehension. For that, we compared the processing of long complex nominal compounds (CNCs) with that of a longer structure composed of a noun followed by prepositional phrases (NPPs). The experimental procedure and items were identical to those of Gamboa et al. (2024), but the participants were L2 speakers of English. We recruited these participants in order to increase the difficulty of the reading task, since Gamboa et al. had suggested that their unclear pattern results may have stemmed from the fact that comprehension tends to be an easier task than production. We focused on two groups of L1 speakers: one whose L1 does not use compounds frequently (Portuguese and Spanish), and one whose L1 does use compounds frequently (German). We predicted, based on the UID hypothesis, that CNCs would lead to more difficulty than NPPs, and that this difficulty would be starker for German speakers than for Portuguese and Spanish speakers.

Our results did partially support the UID, but were not as clear as expected. They indicate the Uniform Information Density hypothesis may not always hold for comprehension. Further research is encouraged to uncover the specific situations in which it does.

Data Availability

The data that support the findings of this study are openly available in OSF at `https://osf.io/cg89w/?view_only=023459fb0f184e6daa07bbd420da6e93` .

Competing Interests

Competing interests: The author(s) declare none.

References

Andrews, S., Veldre, A., & Clarke, I. E. (2020). Measuring lexical quality: The role of spelling ability. *Behavior Research Methods*, *52*(6), 2257–2282. https://doi.org/10.3758/s13428-020-01387-3

Baddeley, A. (2011). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*, 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bartolic, L. (1978). Nominal compounds in technical English. In L. Trimble, M. Trimble, & K. Drobnic (Eds.), *English for specific purposes: Science and technology* (pp. 257–277). Corvallis, OR: Oregon State University.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. (R package version 1.1-18-1) https://doi.org/10.18637/jss.v067.i01

Berg, T. (2012). The cohesiveness of English and German compounds. *The Mental Lexicon*, *7*(1), 1–33. https://doi.org/10.1075/ml.7.1.01ber

Berg, T. (2016). The semantic structure of English and German compounds: Same or different? *Studia Neophilologica*, *88*(2), 148–164. https://doi.org/10.1080/00393274.2015.1135758

Biber, D., & Gray, B. (2011). Grammatical change in the noun phrase: The influence of written language use. *English Language & Linguistics*, *15*(2), 223–250. https://doi.org/10.1017/S1360674311000025

Carrió Pastor, M. L. (2008). English complex noun phrase interpretation by Spanish learners. *Revista Española de Lingüística Aplicada*, *21*, 27–44. Retrieved from `https://dialnet.unirioja.es/servlet/articulo?codigo=2925910`

Carrió Pastor, M. L., & Candel Mora, M. Á. (2013). Variation in the translation patterns of English complex noun phrases into Spanish in a specific domain.

Languages in Contrast, *13*(1), 28–45. https://doi.org/10.1075/lic.13.1.02car

Christianson, K., Luke, S. G., Hussey, E. K., & Wochna, K. L. (2017). Why reread? Evidence from garden-path and local coherence structures. *The Quarterly Journal of Experimental Psychology*, *70*(7), 1380–1405. https://doi.org/10.1080/17470218.2016.1186200

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics*, *27*(1), 3–42. https://doi.org/10.1017/S0142716406060024

Collins, M. X. (2014). Information density and dependency length as complementary cognitive models. *Journal of Psycholinguistic Research*, *43*, 651–681. https://doi.org/10.1007/s10936-013-9273-3

Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering*, *14*(3), 391–403. https://doi.org/10.1081/QEN-120001878

Dan. (2020, August 27). *Get Reading Measures* [Online forum post]. Retrieved 2021-06-21, from `https://www.sr-research.com/support/showthread.php?tid=26`

Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, *40*(4), 296–340. https://doi.org/10.1006/cogp.1999.0730

Forster, K. I., Guerrera, C., & Elliot, L. (2009). The maze task: Measuring forced incremental sentence processing time. *Behavior Research Methods*, *41*(1), 163–171. https://doi.org/10.3758/BRM.41.1.163

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (Third ed.). Thousand Oaks, CA: Sage.

Gamboa, J. C. B., Fernandez, L. B., & Allen, S. E. M. (2024). Investigating the uniform information density hypothesis with complex nominal compounds. *Applied Psycholinguistics*, *45*(2), 322–367. https://doi.org/10.1017/S0142716424000092

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In P. Isabelle, E. Charniak, & D. Lin (Eds.), *Proceedings of the 40th annual meeting of the*

Association for Computational Linguistics (pp. 199–206). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073117

Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5), 389–407. https://doi.org/10.1016/j.tics.2019.02.003

Hopp, H. (2010). Ultimate attainment in L2 inflection: Performance similarities between non-native and native speakers. *Lingua*, *120*(4), 901–931. https://doi.org/10.1016/j.lingua.2009.06.004

Horsella, M., & Pérez, F. (1991). Nominal compounds in chemical English literature: Toward an approach to text typology. *English for Specific Purposes*, *10*(2), 125–138. https://doi.org/10.1016/0889-4906(91)90005-H

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62. https://doi.org/10.1016/j.cogpsych.2010.02.002

Jared, D., & O'Donnell, K. (2017). Skilled adult readers activate the meanings of high-frequency words using phonology: Evidence from eye tracking. *Memory & Cognition*, *45*, 334–346. https://doi.org/10.3758/s13421-016-0661-4

Jiang, N. (2007). Selective integration of linguistic knowledge in adult second language learning. *Language Learning*, *57*(1), 1–33. https://doi.org/10.1111/j.1467-9922.2007.00397.x

Jones, D., Farrand, P., Stuart, G., & Morris, N. (1995). Functional equivalence of verbal and spatial information in serial short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 1008–1018. https://psycnet.apa.org/doi/10.1037/0278-7393.21.4.1008

Jullian, P. (2001). Mental representation of English complex nominals by Spanish speakers. *Onomázein*, *6*, 239–247. https://doi.org/10.7764/onomazein.6.13

Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different? *Linguistic Approaches to Bilingualism*, *4*(2), 257–282.

https://doi.org/10.1075/lab.4.2.05kaa

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H., et al. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. (R package version 3.0.1) https://doi.org/10.18637/jss.v082.i13

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, *44*(2), 325–343. https://doi.org/10.3758/s13428-011-0146-0

Levy, R., & Jaeger, T. F. (2006). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. C. Platt, & T. Hoffman (Eds.), *Proceedings of the 19th International Conference on Neural Information Processing Systems* (pp. 849–856). Cambridge, MA: MIT Press. Retrieved from `https://proceedings.neurips.cc/paper/2006/hash/c6a01432c8138d46ba39957a8250e027-Abstract.html`

McDonald, J. L. (2006). Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners. *Journal of Memory and Language*, *55*(3), 381–401. https://doi.org/10.1016/j.jml.2006.06.006

Meister, C., Pimentel, T., Haller, P., Jäger, L., Cotterell, R., & Levy, R. (2021). Revisiting the Uniform Information Density hypothesis. In M.-F. Moens, X. Huang, L. Specia, & S. W. tau Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 963–980). Stroudsburg, PA: Association for Computational Linguistics. http://doi.org/10.18653/v1/2021.emnlp-main.74

Oliveira, C., Freitas, M. C., Quental, V., dos Santos, C. N., Leme, R. P., & Souza, L. (2006). A set of NP-extraction rules for Portuguese: Defining, learning and pruning. In R. Vieira, P. Quaresma, M. d. G. V. Nunes, N. J. Mamede, C. Oliveira, & M. C. Dias (Eds.), *International Workshop on Computational Processing of the Portuguese Language* (pp. 150–159). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/11751984_16

Olshtain, E. (1981). English nominal compounds and the ESL/EFL reader. In M. Hines & W. Rutherford (Eds.), *On TESOL '81: Selected papers from the fifteenth annual Conference of Teachers of English to Speakers of Other Languages* (pp. 153–168). Washington, DC: TESOL. Retrieved from https://eric.ed.gov/?id=ED223079

Paape, D., & Vasishth, S. (2022). Does conscious rereading lead to targeted regressions in garden-path sentences? Data from a novel stop-and-reread paradigm. *PsyArXiv preprint*. https://doi.org/10.31234/osf.io/vjyfz

Pabón Berbesí, M., & Domínguez, C. L. (2009). Structure and function of the nominal group in English and Spanish in academic texts. In S. Burgess & P. Martín-Martín (Eds.), *English as an additional language in research publication and communication* (pp. 215–235). Berlin: Peter Lang.

Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. In L. Verhoeven, C. Elbro, & P. Reitsma (Eds.), *Precursors of functional literacy* (pp. 189–213). Amsterdam: Benjamins. https://doi.org/10.1075/swll.11.14per

Pickering, M. J., & Traxler, M. J. (1998). Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(4), 940. https://doi.org/10.1037/0278-7393.24.4.940

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria. Retrieved from https://www.r-project.org/ (Version 3.4.0)

Richman, S. (1969). The translation to Spanish of English nouns in juxtaposition. *Hispania*, *52*(3), 426–430. https://doi.org/10.2307/337897

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sikos, L., Greenberg, C., Drenhaus, H., & Crocker, M. W. (2017). Information density of encodings: The role of syntactic variation in comprehension. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 3168–3173).

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2024). afex: Analysis of factorial experiments [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=afex (R package version 1.4-1)

Solso, R. L., & King, J. F. (1976). Frequency and versatility of letters in the English language. *Behavior Research Methods & Instrumentation*, *8*(3), 283–286. https://doi.org/10.3758/BF03201714

Trimble, M. T., & Trimble, L. (1977). The development of EFL materials for occupational English. In H. L. B. Moody & J. D. Moore (Eds.), *English for Specific Purposes: An International Seminar* (pp. 57–70). Bogotá, Colombia: The British Council.

van Helmond, K., & van Vugt, M. (1985). On the transferability of nominal compounds. *Interlanguage Studies Bulletin*, 5–34. Retrieved from https://www.jstor.org/stable/43135308

von der Malsburg, T., & Angele, B. (2017). False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language*, *94*, 119–133. https://doi.org/10.1016/j.jml.2016.10.003

Footnotes

[1] While this conversion of the Morse code into sequences of 0, 1 and 2 is possible, it is certainly not how Morse code is implemented in real applications. This is because most symbols 0 would be unnecessary or redundant: they appear only as a separator between dots and dashes. This would reduce the amount of information that could be sent through time, and therefore is an example of a code that is not efficient (see below). We use this example here only for illustration purposes.

[2] Of course, this is only true if we disregard the context in which these letters occur. For example, given a message containing the sequence of letters "*m, y, f, a, v, o, r, i, t, e, m, u, s, i, c, g, e, n, r, e, i, s, j, a, z*", one would quite strongly expect that the next letter is a *z*, and therefore the amount of information transmitted by *z* would be low.

[3] As has been previously done in the literature (see, e.g. Levy & Jaeger, 2006; Genzel & Charniak, 2002)

[4] Note that the experiment reported by Sikos et al. (2017) used structures that look similar to CNCs (*carefully written essay*) and NPPs (*essay that was carefully written*). However, their predictions were about the processing time for the *head noun* and not about the structures as a whole. The head noun,

being the last word of CNCs and first word of NPPs, is likely less informative in CNCs than NPPs, even if CNCs are more informative *as a whole*. Therefore, their predictions are inverted in comparison to those of Gamboa et al. (2024) and the experiments reported here.

[5] All items, data and code used in this study can be found in the associated OSF page `https://osf.io/cg89w/?view_only=023459fb0f184e6daa07bbd420da6e93` .

[6] Gamboa et al. (2024) also excluded trials that contained five or fewer fixations. This was not the case for any trial in our data.

[7] These measures were calculated using the Get Reading Measures tool (Dan, 2020).

[8] We used the following R formula for all models, including the one used for the regression count (see below): `DV ~ OPT + DST + length*type*language + (1 + length*type | subject) + (1 + DST + OPT + type*length*language | item)` . In cases where the models did not converge, we removed the model covariance terms using the afex package (Singmann, Bolker, Westfall, Aust, & Ben-Shachar, 2024). For that, we used their `lmer_alt` function and replaced every "`|`" of the formula above with "`||`".

[9] We used the following formula in R for all models, including the one used for the regression count (see below): `DV ~ OPT + Spelling + length*type*language + VerbalWM + VisualWM + (1 + length*type | subject) + (1 + VerbalWM + VisualWM + type*length*language | item)` . Different from Experiment 1, no model presented convergence problems, so we did not need to remove covariance terms for any model.

[10] The Variance Inflation Factor measures how well a given variable can be estimated based on the other variables of the model. Typically, one should avoid VIF scores above 10 (Craney & Surles, 2002).

[11] Indeed, for Experiment 2, we do not believe that our non-significant findings are due to poorly fit models resulting from the significantly correlating proficiency variables, since the VIF values were all very low.

[12] Still, note that the majority of the participants in these groups were recruited among international Master students enrolled in English-language programs. We presume most of them did not reguarly read in German or use academic German.

Table 1

Example sentences for each experimental condition in Experiment 1. Adapted from Gamboa et al. (2024)

Condition	**Preamble**	**Critical Region**	**Region of Interest 1 (ROI1)**	**Region of Interest 2 (ROI2)**	**Wrap Up**
4-CNC	In present times,	health insurance economy effects	are researched	by the analysts	of financial institutions
4-NPP	In present times,	effects of the economy on health insurance	are researched	by the analysts	of financial institutions
6-CNC	In current times,	United States factory employee insurance costs	are decreased	by the changes	in union policies
6-NPP	In current times,	the cost of insurance for factory employees in the United States	are decreased	by the changes	in union policies
Condition	**Comprehension Questions**				
4-CNC/NPP	Financial institution analysts research health insurance X: yearly costs M: economy effects				
6-CNC/NPP	Union policy changes decrease employee insurance costs in X: the United States M: France				

Table 2

Participant accuracy for Experiment 1, along with the number of discarded trials.

Language	Type	Length	Mean Accuracy	SD	# Trials Correct	Total Trials
German	NP	4	0.957	0.074	201	210
German	NP	6	0.924	0.117	194	210
German	PP	4	0.967	0.068	203	210
German	PP	6	0.967	0.079	203	210
P/S	NP	4	0.934	0.102	185	198
P/S	NP	6	0.970	0.077	192	198
P/S	PP	4	0.949	0.088	188	198
P/S	PP	6	0.980	0.055	194	198
TOTAL					1560 (95.59%)	1632
DISCARDED TRIALS					Total incorrect	72
					Too noisy before correction	6
					Too noisy after correction	3
					Final dataset analysed	1551

Table 3

Trials trimmed prior to analysis of duration measures in Experiment 1.

	Region of Interest 1		**Region of Interest 2**	
	Discarded	Kept	Discarded	Kept
FPD	345 (22.24%)	1206	187 (12.06%)	1364
TD	148 (9.54%)	1403	157 (10.12%)	1394

Table 4

Mean and standard deviation of Experiment 1 measures (in ms for duration measures, and in number of occurrences for the regression). Regression values should be taken with care, since the distribution is not normal (see Figure 3).

Length	**Structure**	**Language**	**FFD ROI1**	**TD ROI1**	**FFD ROI2**	**TD ROI2**	**Reg2CR**
3 Nouns	CNC	German	531.9 (200.1)	1002.2 (462.7)	481.2 (206.9)	913.4 (438.5)	1.9 (2.0)
		Spanish	503.6 (197.9)	937.7 (441.6)	426.3 (184.0)	905.1 (439.6)	2.9 (2.9)
	NPP	German	537.3 (186.2)	870.9 (426.0)	518.9 (232.4)	949.1 (465.6)	2.8 (2.5)
		Spanish	486.1 (170.1)	819.4 (430.6)	440.4 (206.3)	907.5 (464.4)	4.1 (3.6)
4 Nouns	CNC	German	564.2 (206.7)	977.4 (462.2)	495.1 (236.6)	948.5 (448.6)	2.8 (2.7)
		Spanish	545.8 (189.4)	917.3 (437.1)	446.7 (209.7)	913.3 (418.6)	4.5 (4.1)
	NPP	German	527.5 (184.4)	923.7 (463.5)	442.9 (225.3)	891.4 (455.3)	4.1 (3.7)
		Spanish	476.7 (167.1)	816.2 (404.3)	436.5 (201.4)	843.5 (441.2)	5.8 (5.5)

Table 5

An example critical item. Adapted from Gamboa et al. (2024)

Condition	**Preamble**	**Critical Region**	**Region of Interest 1**	**Region of Interest 2**	**Wrap Up**
3-NC	In present times,	the inflation constraint action	is implemented	by the board	of the national bank
4-NC	In present times,	the currency inflation constraint action	is implemented	by the board	of the national bank
3-PP	In present times,	the action for the constraint of inflation	is implemented	by the board	of the national bank
4-PP	In present times,	the action for the constraint of inflation of the currency	is implemented	by the board	of the national bank
Comprehension question	The national bank board implements X: inflation constraint actions M: deflation constraint actions				

Table 6

Participant accuracy for Experiment 2, along with the number of discarded trials.

Language	Type	Length	Mean Accuracy	SD	# Trials Correct	Total Trials
German	NP	3	0.925	0.112	162	175
German	NP	4	0.863	0.155	151	175
German	PP	3	0.910	0.123	161	177
German	PP	4	0.875	0.109	153	175
Spanish	NP	3	0.962	0.068	182	189
Spanish	NP	4	0.903	0.133	171	189
Spanish	PP	3	0.941	0.096	177	188
Spanish	PP	4	0.912	0.113	173	190
TOTAL					1330 (91.22%)	1458
DISCARDED TRIALS					Total incorrect	128
					Too noisy before correction	17
					Too noisy after correction	8
					Final dataset analysed	1305

Table 7

Trials trimmed prior to analysis of duration measures in Experiment 2.

	Region of Interest 1		**Region of Interest 2**	
	Discarded	Kept	Discarded	Kept
FPD	235 (18.01%)	1070	137 (10.50%)	1168
TD	138 (10.57%)	1167	134 (10.27%)	1171

Table 8

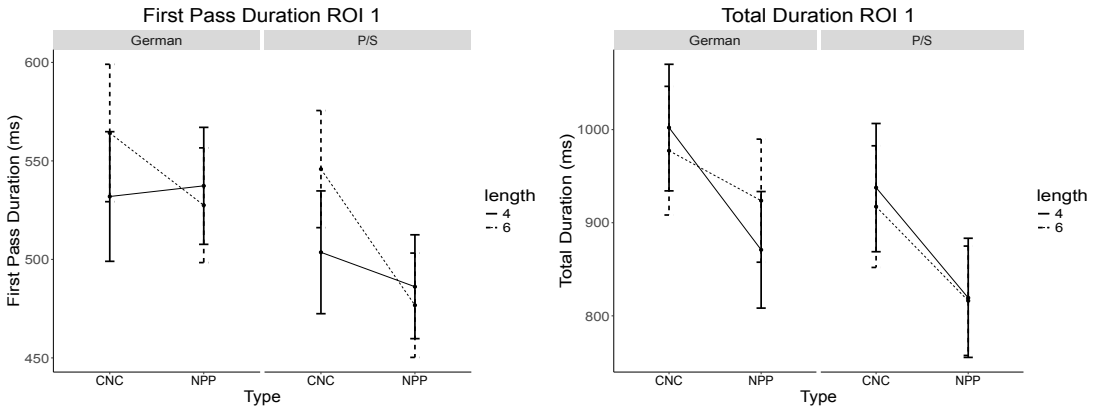
Mean and standard deviation of Experiment 2 measures (in ms for duration measures, and in number of occurrences for the regression). Regression values should be taken with care, since the distribution is not normal (see Figure 6).

Length	**Structure**	**Language**	**FFD ROI1**	**TD ROI1**	**FFD ROI2**	**TD ROI2**	**Reg2CR**
3 Nouns	CNC	German	576.6 (215.6)	896.2 (419.9)	535.4 (195.4)	915.9 (404.9)	1.5 (1.6)
		Spanish	482.0 (198.5)	885.5 (439.4)	457.4 (212.3)	854.0 (440.5)	1.7 (1.9)
	NPP	German	532.2 (194.4)	906.7 (425.6)	513.4 (227.7)	936.2 (398.7)	2.7 (2.6)
		Spanish	472.3 (188.7)	815.2 (454.1)	459.7 (220.6)	855.7 (417.8)	3.2 (2.8)
4 Nouns	CNC	German	527.0 (168.5)	844.0 (418.1)	499.8 (221.8)	885.2 (368.8)	1.5 (1.8)
		Spanish	495.8 (197.7)	841.1 (445.7)	460.2 (211.9)	823.9 (392.8)	2.3 (2.5)
	NPP	German	532.4 (207.7)	843.6 (427.9)	464.5 (219.0)	897.5 (436.0)	3.5 (3.6)
		Spanish	459.7 (184.6)	813.8 (409.1)	451.1 (202.8)	815.3 (416.0)	4.3 (3.8)

Table 9

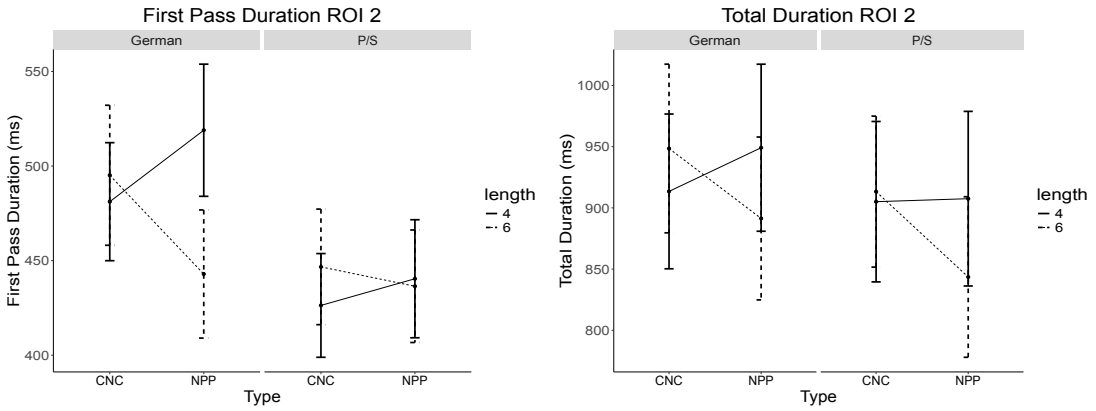
Summary of results of this study, along with the L1 results from Gamboa et al. (2024).

Experiment 1				
	L1		**L2**	
	Finding	Predicted?	Finding	Predicted?
Critical	Reg2CR: NPP > CNC	No	Reg2CR: NPP > CNC	No
Region	Reg2CR: 6 > 4	Yes	Reg2CR: 6 > 4	Yes
			Reg2CR: P/S > DE	No
ROI 1	↑DST → ↓TD	Yes	TD: CNC > NPP	Yes
			TD: DE > P/S	No
ROI 2	↑DST → ↓TD	Yes		
Experiment 2				
	L1		**L2**	
	Finding	Predicted?	Finding	Predicted?
Critical	Reg2CR: NPP > CNC	No	Reg2CR: NPP > CNC	No
Region	Reg2CR: 4 > 3	Yes	Reg2CR: 4 > 3	Yes
			↑Spelling → ↑Reg2CR	No
ROI 1			↑Spelling → ↓FPD	Yes
			↑VerbalWM → ↓TD	Yes
ROI 2	FPD: CNC > NPP	Yes	↑Spelling → ↓FPD	Yes
			↑VerbalWM → ↓TD	Yes



(a) (b)

Figure 1. Experiment 1: First Pass Durations and Total Durations in Region of Interest 1. The error bars are 95% confidence intervals.



(a) (b)

Figure 2. Experiment 1: First Pass Durations and Total Durations in Region of Interest 2. The error bars are 95% confidence intervals.

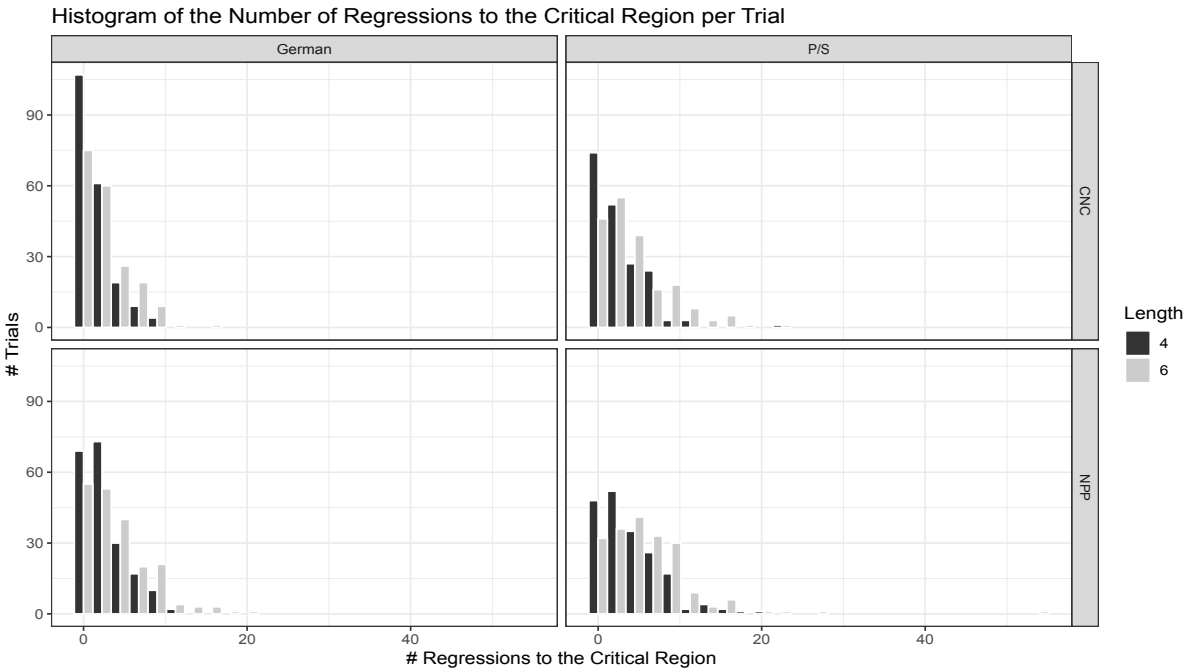
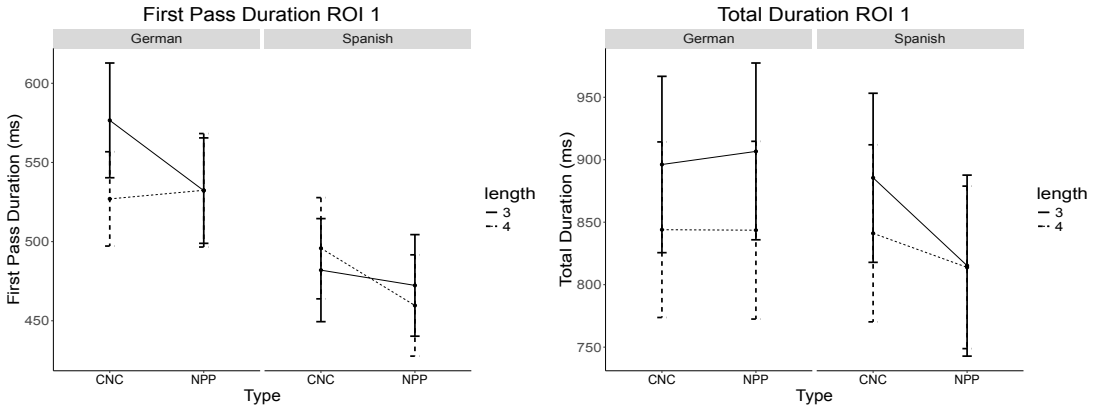
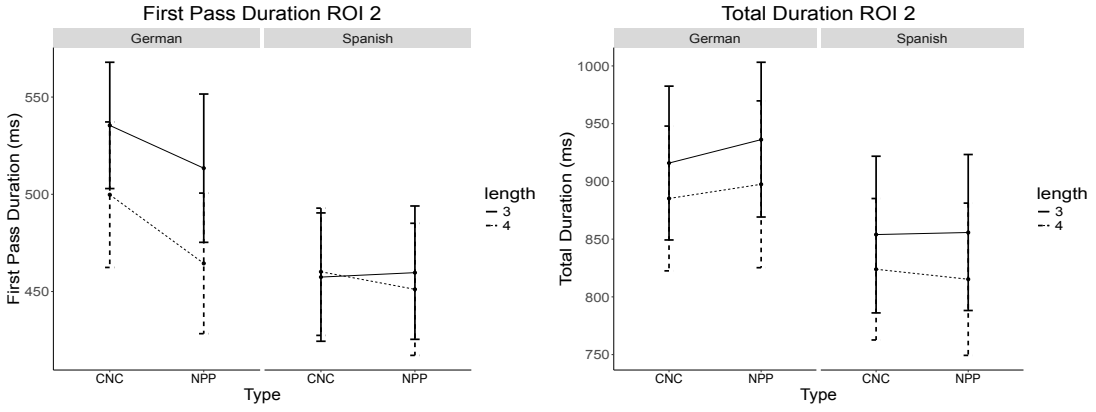


Figure 3. Experiment 1: Regressions towards the critical region.



(a) (b)

Figure 4. Experiment 2: First Pass Durations and Total Durations in Region of Interest 1. The error bars are 95% confidence intervals.



(a) (b)

Figure 5. Experiment 2: First Pass Durations and Total Durations in Region of Interest 2. The error bars are 95% confidence intervals.

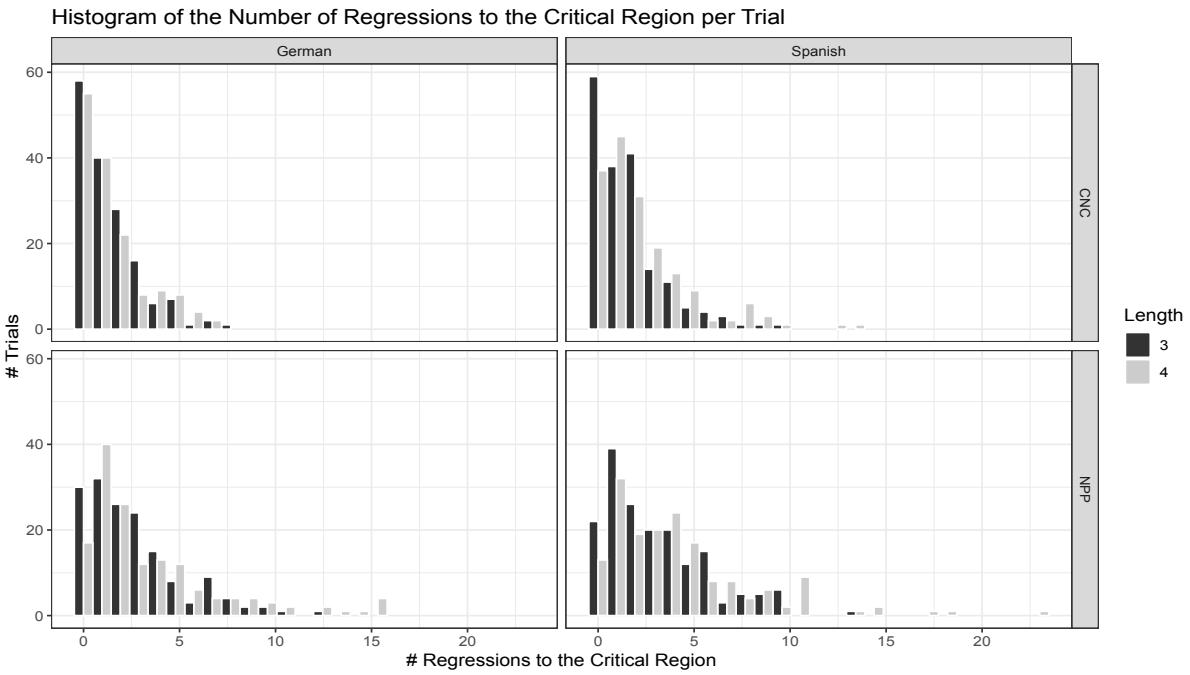


Figure 6. Experiment 2: Regressions towards the critical region.

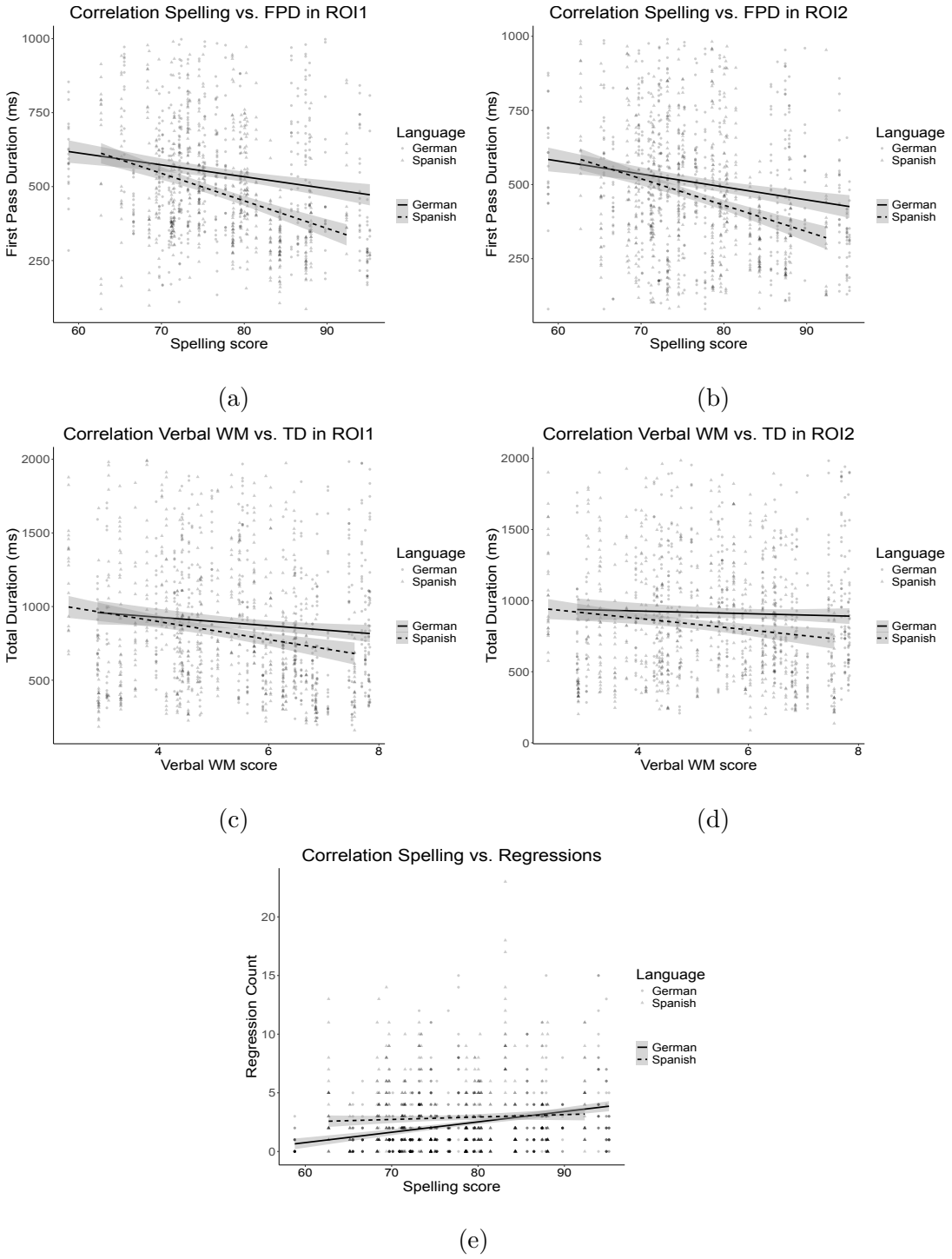


Figure 7. The relation between the covariates and the reading measure they were found to predict. Data points are made transparent for a sense of density. Gray shading indicates 95% confidence intervals for the regression line.

Appendix A

Experiment 1 Full Model Results

Significant (p<.01) are highlighted in bold.

Table A1

Experiment 1: First Pass Duration – Region of Interest 1.

	b	SE	t	p
Intercept	**550.102**	**10.828**	**50.802**	**<0.001**
OPT	-4.757	9.290	-0.512	0.609
DST	-8.018	8.675	-0.924	0.355
length	-8.716	17.027	-0.512	0.609
type	20.039	14.553	1.377	0.169
language	37.150	19.607	1.895	0.058
length:type	-45.592	28.330	-1.609	0.108
length:language	2.350	27.682	0.085	0.932
type:language	-19.050	24.912	-0.765	0.444
length:type:language	19.093	48.005	0.398	0.691

Table A2

Experiment 1: First Pass Duration – Region of Interest 2.

	b	SE	t	p
Intercept	**475.376**	**10.978**	**43.302**	**<0.001**
OPT	1.204	10.162	0.118	0.906
DST	9.499	9.828	0.966	0.334
length	15.857	18.167	0.873	0.383
type	-1.391	15.528	-0.090	0.929
language	38.477	19.985	1.925	0.054
length:type	-52.206	29.191	-1.788	0.074
length:language	39.612	30.848	1.284	0.199
type:language	-0.739	27.634	-0.027	0.979
length:type:language	-28.683	51.502	-0.557	0.578

Table A3

Experiment 1: Total Duration – Region of Interest 1.

	b	SE	t	p
Intercept	**1007.120**	**26.530**	**37.963**	**<0.001**
OPT	-52.070	24.690	-2.109	0.035
DST	18.450	22.730	0.812	0.417
length	18.020	40.230	0.448	0.654
type	102.850	26.870	3.828	**0.000**
language	115.660	44.840	2.579	**0.010**
length:type	23.890	55.830	0.428	0.669
length:language	-3.770	54.450	-0.069	0.945
type:language	-18.640	46.680	-0.399	0.690
length:type:language	9.000	98.170	0.092	0.927

Table A4

Experiment 1: Total Duration – Region of Interest 2.

	b	SE	t	p
Intercept	**1011.716**	**27.143**	**37.273**	**<0.001**
OPT	-17.140	22.718	-0.754	0.451
DST	1.813	23.012	0.079	0.937
length	-1.739	40.522	-0.043	0.966
type	15.990	29.471	0.543	0.587
language	68.639	46.405	1.479	0.139
length:type	-92.129	58.465	-1.576	0.115
length:language	-21.287	55.596	-0.383	0.702
type:language	-16.091	50.550	-0.318	0.750
length:type:language	10.509	100.115	0.105	0.916

Table A5

Experiment 1: Regressions onto the Critical Region.

	b	SE	z	p
Intercept	**1.039**	**0.075**	**13.789**	**<0.001**
OPT	0.071	0.073	0.969	0.332
DST	0.034	0.073	0.463	0.643
length	-0.384	0.063	-6.138	**0.000**
type	-0.368	0.064	-5.802	**0.000**
language	-0.453	0.145	-3.130	**0.002**
length:type	-0.085	0.122	-0.694	0.488
length:language	0.030	0.074	0.401	0.688
type:language	-0.070	0.096	-0.727	0.467
length:type:language	0.023	0.180	0.126	0.900

Appendix B

Experiment 2 Full Model Results

Significant (p<.01) are highlighted in bold.

Table B1

Experiment 2: First Pass Duration – Region of Interest 1.

	b	SE	t	p
Intercept	**547.800**	**11.190**	**48.951**	**<0.001**
OPT	12.640	13.060	0.967	0.333
Spelling	-34.740	11.290	-3.077	**0.002**
length	11.170	11.040	1.012	0.311
type	22.090	14.350	1.540	0.124
language	47.050	23.280	2.021	0.043
VerbalWM	-28.800	13.190	-2.184	0.029
VisualWM	22.100	11.480	1.925	0.054
length:type	38.390	26.160	1.467	0.142
length:language	15.410	22.140	0.696	0.486
type:language	12.390	26.040	0.476	0.634
length:type:language	1.890	50.080	0.038	0.970

Table B2

Experiment 2: First Pass Duration – Region of Interest 2.

	b	SE	t	p
Intercept	**501.677**	**11.585**	**43.305**	**<0.001**
OPT	19.800	13.527	1.464	0.143
Spelling	-61.173	11.598	-5.274	**0.000**
length	24.867	13.189	1.885	0.059
type	16.224	12.725	1.275	0.202
language	27.330	24.012	1.138	0.255
VerbalWM	-10.288	13.160	-0.782	0.434
VisualWM	9.164	11.731	0.781	0.435
length:type	2.605	23.531	0.111	0.912
length:language	41.986	23.562	1.782	0.075
type:language	30.895	27.654	1.117	0.264
length:type:language	-8.981	51.801	-0.173	0.862

Table B3

Experiment 2: Total Duration – Region of Interest 1.

	b	SE	t	p
Intercept	**842.039**	**26.662**	**31.582**	**<0.001**
OPT	8.894	30.495	0.292	0.771
Spelling	-44.648	26.341	-1.695	0.090
length	27.952	24.740	1.130	0.259
type	22.250	29.461	0.755	0.450
language	105.229	55.852	1.884	0.060
VerbalWM	-92.284	29.326	-3.147	**0.002**
VisualWM	59.210	28.539	2.075	0.038
length:type	34.844	56.282	0.619	0.536
length:language	76.513	47.659	1.605	0.108
type:language	-30.295	57.414	-0.528	0.598
length:type:language	-29.524	112.884	-0.262	0.794

Table B4

Experiment 2: Total Duration – Region of Interest 2.

	b	SE	t	p
Intercept	**941.775**	**24.606**	**38.274**	**<0.001**
OPT	33.733	27.055	1.247	0.212
Spelling	-18.757	23.888	-0.785	0.432
length	47.104	23.943	1.967	0.049
type	-8.722	27.705	-0.315	0.753
language	92.233	50.968	1.810	0.070
VerbalWM	-85.648	26.996	-3.173	**0.002**
VisualWM	51.115	24.858	2.056	0.040
length:type	37.226	55.579	0.670	0.503
length:language	51.620	40.730	1.267	0.205
type:language	-4.554	52.678	-0.086	0.931
length:type:language	-99.717	99.854	-0.999	0.318

Table B5

Experiment 2: Regressions onto the Critical Region.

	b	SE	z	p
Intercept	**0.626**	**0.086**	**7.249**	**<0.001**
OPT	-0.057	0.108	-0.527	0.598
Spelling	0.287	0.094	3.035	**0.002**
length	-0.196	0.061	-3.209	**0.001**
type	-0.747	0.075	-9.982	**<0.001**
language	-0.184	0.201	-0.920	0.358
VerbalWM	-0.160	0.106	-1.509	0.131
VisualWM	0.115	0.097	1.185	0.236
length:type	0.094	0.101	0.933	0.351
length:language	0.169	0.120	1.406	0.160
type:language	-0.107	0.122	-0.875	0.382
length:type:language	0.156	0.216	0.723	0.470

Appendix C

Experiment 2 Covariate Results

[h]

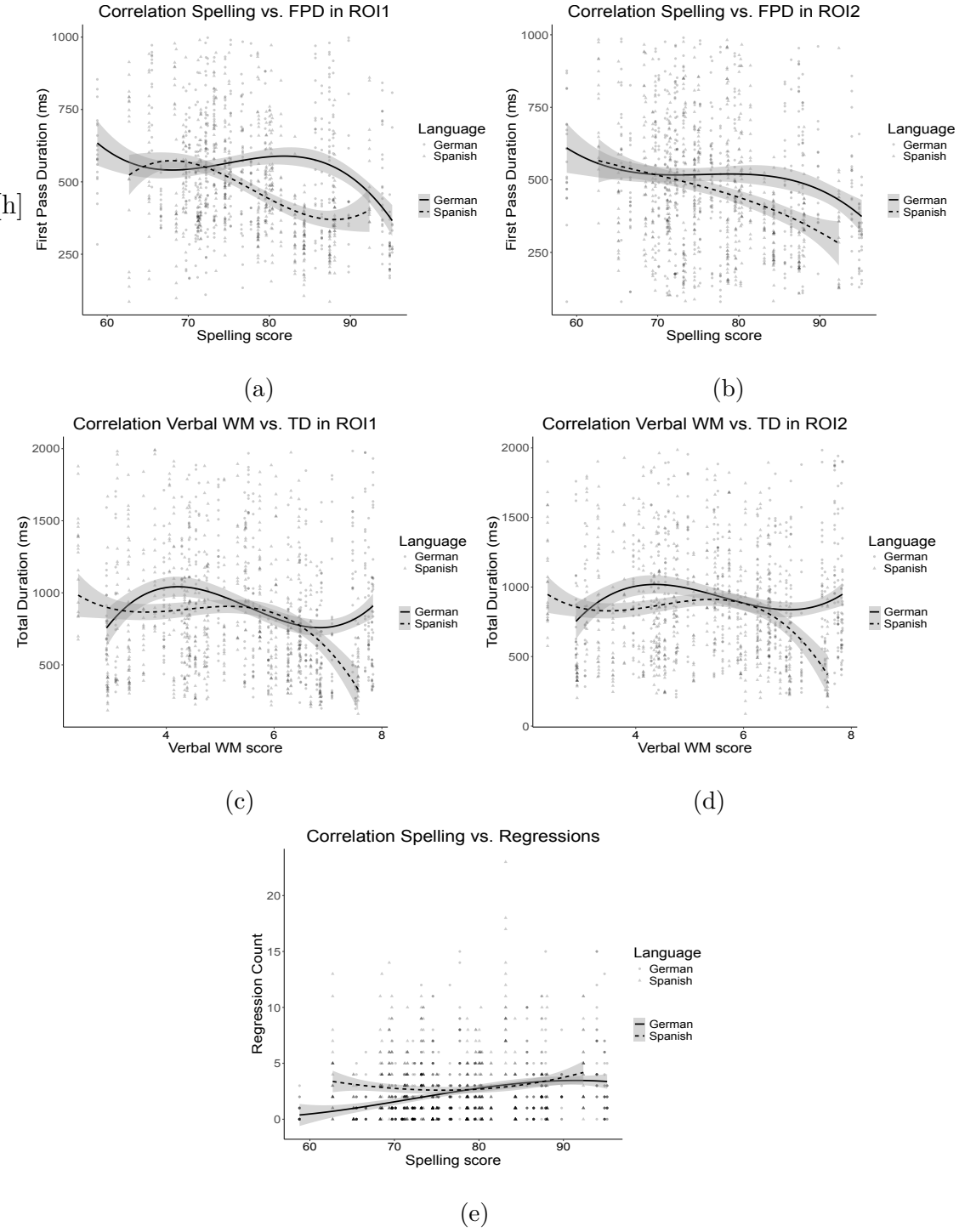
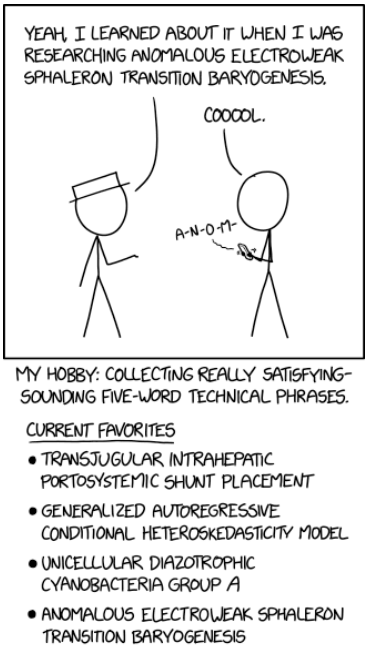


Figure C1. The graphs of Figure C1 with cubic splines. Gray shading indicates 95% confidence intervals.

6 Publications on CNC use



Five Word Jargon

Mouseover text: My other (much harder) hobby is trying to engineer situations where I have an excuse to use more than one of them in short succession.

Munroe, 2020a

Contents

6.1 How CNCs are set up in scientific papers 153
6.2 The role of context in the perceived difficulty of CNCs 154

In this chapter, the publications associated with the second research question are reprinted. Once again, the publications are inserted as they were published and/or submitted, and the page numbers may not align with the page numbers of this dissertation.

The two publications are described in the following sections.

6.1 How CNCs are set up in scientific papers

The first publication in this chapter is:

Gamboa, J., Braun, K., Järvikivi, J. & Allen, S. E. M. (2024). The distributional properties of long nominal compounds in scientific articles: an investigation based on the uniform information density hypothesis. *Corpus Linguistics and Linguistic Theory, 21*(1), 137–171. https://doi.org/10.1515/cllt-2023-0028

The authors contributed to the publication in the following ways:

- **Conceptualization:** Allen, Braun, Gamboa, Järvikivi;
- **Corpus construction:** Braun, Gamboa;
- **Methodology:** Braun, Gamboa;

- **Software:** Braun, Gamboa;
- **Visualization:** Gamboa;
- **Quantitative analysis:** Gamboa;
- **Qualitative analysis:** Braun, Gamboa;
- **Interpretation of results:** Allen, Gamboa;
- **Writing (first draft):** Gamboa;
- **Writing (review and editing):** Allen, Braun, Gamboa, Järvikivi;
- **Supervision:** Allen, Järvikivi;
- **Funding acquisition:** Allen;
- All authors approved the submitted version of the manuscript.

6.2 The role of context in the perceived difficulty of CNCs

The second publication in this chapter is:

Gamboa, J., Gvozdeva, D., Ito, F., Järvikivi, J. & Allen, S. E. M. (2025). *How does the context preceding a long nominal compound influence its comprehension difficulty?* [Manuscript submitted for publication]. Psycholinguistics and Language Development Research Group, RPTU University of Kaiserslauter-Landau.

The authors contributed to the publication in the following ways:

- **Conceptualization:** Allen, Gamboa, Järvikivi;
- **Methodology:** Gamboa, Gvozdeva, Ito;
- **Software:** Gamboa, Gvozdeva, Ito;
- **Visualization:** Gamboa, Gvozdeva, Ito;
- **Data collection:** Gamboa, Gvozdeva, Ito;
- **Data curation:** Gamboa, Gvozdeva, Ito;
- **Data analysis:** Gamboa;
- **Interpretation of results:** Allen, Gamboa;
- **Writing (first draft):** Gamboa;
- **Writing (review and editing):** Allen, Gamboa, Gvozdeva, Ito, Järvikivi;
- **Supervision:** Allen, Järvikivi;
- **Funding acquisition:** Allen;
- All authors approved the submitted version of the manuscript.

Article

John Gamboa*, Kristina Braun, Juhani Järvikivi and Shanley E. M. Allen

The distributional properties of long nominal compounds in scientific articles: an investigation based on the uniform information density hypothesis

https://doi.org/10.1515/cllt-2023-0028
Received March 13, 2023; accepted March 1, 2024; published online April 17, 2024

Abstract: Nominal compounds are a structure commonly used in scientific texts. Despite their commonality, very little is known about how they are distributed in scientific articles. Based on the Uniform Information Density hypothesis, which states that speakers communicate information at a constant rate, avoiding peaks and troughs of information transmission, we predict that nominal compounds should cluster toward the end of scientific texts, be preceded by supporting text that facilitates their understanding, and be repeated often after their first use. In this paper, we examine these predictions through a quantitative and a qualitative analysis of a corpus of scientific papers from the fields of Biology, Economics and Linguistics. While our investigation did not reveal definitive findings for the first and third predictions above, it did produce supporting evidence in favor of our second prediction, thus advancing our understanding of NC use and the choices speakers make when transmitting information.

Keywords: nominal compounds; uniform information density hypothesis; scientific writing

***Corresponding author: John Gamboa**, University of Kaiserslautern-Landau, Kaiserslautern, Rheinland-Pfalz, Germany, E-mail: gamboa@rptu.de. https://orcid.org/0000-0003-2430-9902
Kristina Braun, DB Systel GmbH, Frankfurt, Germany, E-mail: christy.kolesova@gmail.com
Juhani Järvikivi, Department of Linguistics, University of Alberta, Alberta, Canada, E-mail: jarvikivi@ualberta.ca. https://orcid.org/0000-0002-3941-2905
Shanley E. M. Allen, University of Kaiserslautern-Landau, Kaiserslautern, Rheinland-Pfalz, Germany, E-mail: allen@rptu.de. https://orcid.org/0000-0002-5421-6750

1 Introduction

All languages of the world allow for some form of compounding (Dressler 2006). In English, the result of this process can be a word (e.g., *outstanding*, *snowman*, *strawberry*, *highlight*), or a more complex structure composed of multiple words (e.g., *blood moon*, *health care provider*, *dopamine production suppressor protein*). This process of word formation is so pervasive that it has been referred to as the "universally fundamental word formation process" (itself a compound), offering "the easiest and most effective way to create and transfer new meanings" (Libben 2006). In English, the vast majority of compounds are nouns (Algeo and Algeo 1991: 7). These are ubiquitous in everyday language, comprising 2.6 % of the British National Corpus and 3.9 % of the Reuters corpus (Baldwin and Tanaka 2004).

In this paper, we are concerned with the sort of compounding that leads to the formation of complex structures composed of multiple words. In particular, we turn our focus to structures we refer to as *nominal compounds* (NC), multiword structures that, as a whole are categorized as nouns. For concreteness, we will focus on examples such as *water waste* or *addictive substance* (both of which, taken as a whole, form a noun) and not on *resource poor* (an adjective), or *ambulance-chase* (a verb).

On the surface, such NCs are typically composed of a head noun and one or more modifiers. For example, the NC *health care* is composed of the head noun *care* and the modifier *health*. On a deeper level, NCs present a hierarchical structure. For example, *health care* could be reused recursively as a single unit in a subsequent compounding process to construct more complicated structures such as *health care provider* (in which it is used as a modifier) or *geriatric health care* (in which it is used as a head noun).[1] In addition, the way in which the NC words are linked can vary widely. For instance, *olive oil* can be interpreted as an "oil FROM olives", but *baby oil* is typically an "oil FOR babies" and an *olive tree* is a "tree THAT HAS olives". While the exact set of possible linking relations between the NC words has been a topic of debate, a number of studies have demonstrated that they have an effect on processing (see, e.g., Gagné and Shoben 1997; Levi 1978; Spalding et al. 2010).

NCs are particularly frequent in scientific texts (Bhatia 1992). In this register, between 9 % and 16 % percent of all words are part of an NC, and NCs tend to be longer than in everyday English (Salager 1984). Biber and Gray (2011), in an analysis of NC use since the 18th century in different English registers, found that their

1 Less often, the head noun may actually precede the modifiers (e.g., *vitamin C*, *attorney general*). Despite their order being reversed, they do follow the same rules as the more common NCs when reused to build other structures. For example, *vitamin C* may be used as a modifier (*vitamin C deficiency*) or as a head noun (*calcium-regulating vitamin C*). In this paper, we ignore these cases.

prevalence in the scientific register increased 10-fold between 1875 and 2005. They also found that their complexity increased, with three-word compounds being initially uncommon but becoming common by the 1950s. They attributed this increase in complexity to a "principle of economy" inherent to these registers. This conclusion echoes those of Montero (1996), who additionally suggested that NC use is the result of a "desire for novelty"; and of Salager (1984: 142), for whom an NC is "a new concept for which the language code has no name … crystallized into a fixed expression owing [*sic*] a scientific meaning which the individual constituents do not have". According to Salager's analysis, this new concept, once used for the first time, can be reused or referred to as an entity that the writers know the readers can understand.

In this paper, we refer to longer NCs (composed of three or more words) as *complex* nominal compounds (CNC). CNCs are sometimes considered hard to understand, and are typically discouraged in "good writing" guidelines (e.g., Tobin 2002). Indeed, a number of studies have shown that they may lead to difficulty in some circumstances. For example, individuals have been shown to be unable to identify the head of a given CNC (Geer et al. 1972; Limaye and Pompian 1991), and the CNCs themselves are ambiguous sometimes (cf. Montero 1996): as noted by Kvam (1990), a *kitchen towel rack* may be a rack for kitchen towels or a towel rack in the kitchen. In addition, L2 speakers (who are common producers and consumers of scientific literature) often translate CNCs in inconsistent ways (Carrió Pastor 2008; Carrió Pastor and Candel Mora 2013), and are sometimes unable to recover the implicit semantic links between the NC words (Horsella and Pérez 1991).

1.1 Information density

Since the beginning of the century, an increasing number of studies have suggested that language use is efficient in an information theoretic sense (see, e.g., Genzel and Charniak 2002, 2003; Hale 2001). Communication, from the point of view of Information Theory (see Shannon 1948), is a transfer of symbols between a transmitter and a receiver through a communication channel, which may be noisy (i.e., some symbols may either arrive to the receiver corrupted or not arrive at all, and the receiver may receive symbols never sent by the transmitter). Under this framework, the transmitter could be a speaker or writer, the receiver could be an interlocutor or a reader, and the communication channel could be the air or paper/screen. The system, therefore, has two goals. First, communication should be performed *reliably*, i.e., all messages and only the messages sent by the transmitter should arrive to the receiver, and they should arrive correctly. Second, communication should be performed *efficiently*: the system should use the minimum amount of resources

possible. Shannon showed that reliable communication using minimum resources is achieved with information being transmitted at a constant rate, the so-called *capacity* of the channel. If the capacity is exceeded, then communication is still possible, but errors are more likely to occur.

What exactly constitutes information is the topic of considerable debate (see Floridi 2009 for a brief review). However, from an information theoretic perspective, the *amount* of information conveyed by a symbol is proportional to how unexpected the symbol is. That is, if a symbol *a* is expected, then it conveys little information; if *a* is surprising, then it conveys a lot of information. This "expectation" can be modulated by any number of factors. For example, if half of the times *a* is transmitted it is preceded by *b*, then the appearance of *b* increases the expectation of seeing *a*, therefore decreasing the information conveyed by *a*.

Implicit in these models is the idea that both transmitter and receiver have perfect knowledge of (or at least agree on) the probabilities of all symbols that can pass through the communication channel. In the case of telegraphs communicating English letters, this allowed for the development of encoding systems (e.g., the Morse code) that were efficient by taking into account the frequency of each letter (e.g., in the Morse code, the encoding of the letter *e* – the most common letter in English – is much shorter than that of the letter *z* – the least frequent letter in English; see Solso and King 1976).

When applied to Psycholinguistics, models based on Information Theory have been quite successful in explaining data at several linguistic levels (e.g., Benjamin and Schmidtke 2023; Frank and Jaeger 2008; Levy and Jaeger 2006; Maurits et al. 2010; Schmidtke et al. 2016). For example, speakers often omit the complementizer *that* when a new clause is expected (i.e., when the *that* is unsurprising), allowing for a more efficient transmission of information (Levy and Jaeger 2006). In addition, the processing of 2-word compounds written together (e.g., *snowball*, *newsroom*) has been shown to be affected by the information associated with the relation linking its two constituents. In a lexical decision task with existing compounds (Schmidtke et al. 2016) and in a self-paced reading task (Benjamin and Schmidtke 2023) with both lexicalized and novel compounds, NCs that could be linked by a large number of competing relations (the relations' probability distribution has a high average[2] amount of information) needed longer to be processed than NCs that are more decidedly linked by one or just a few options (the typical choice of relation is unsurprising). For example, NCs such as *newsroom*, which have been strongly interpreted as "room FOR news", were processed faster than NCs such as *floodlight*, which have been variously interpreted as (among others) "light FROM flood", "flood IS light", and "light DURING flood".

2 This average information amount is typically referred to as *Entropy*.

Given the successful application of Information Theory models on Psycholinguistics, Jaeger (2010: 24) suggested that "language production at *all levels* of linguistic representation is organized to be communicatively efficient" (emphasis added). Thus, he proposed the Uniform Information Density hypothesis (UIDh), stating that "speakers prefer utterances that distribute information uniformly across the signal (information density)" (p. 25). When it is not distributed uniformly (e.g., a peak of information is transmitted at once), the capacity of the communication channel may be exceeded, potentially causing comprehension difficulty.

In the scientific register, longer nominal compounds are arguably dense packages of information, and therefore are good candidates for structures that lead to comprehension difficulty. As a simplified example, consider the CNC *waste water treatment facility*. The whole CNC is composed of four words, and therefore each word transmits on average 25 % of its information. Now consider an alternative structure which conveys the same meaning through the use of prepositions: *facility for the treatment of water from waste* (8 words). If we consider the information contained in both structures to be roughly the same, then each word in the second structure carries on average half of the information contained in each words of the CNC, i.e., the information is better "spread" through the linguistic signal.

It is easy to see how CNCs such as *start arm barrier*, *listener network rank* or *reliable stop signal reaction time*, when considered in isolation, make little sense and could lead to comprehension difficulty. Following Jaeger's words, *speakers* should *prefer utterances that distribute information* more *uniformly across the signal* – for example, by using prepositional phrases. Surprisingly, however, as mentioned above, scientific articles do contain many CNCs. The aforementioned CNCs are from real articles: they were the chosen form! What is going on here?

Applying information theoretic models to natural language communication is difficult because the probabilities of the words are not only unknown, but also depend on extralinguistic information that is often not available to the system. For example, the word *program* is likely to have different meanings (and probabilities) depending on whether it is used by a computer programmer talking to their peers or a musician preparing for a concert. If neither the transmitter nor the receiver has perfect information about the probabilities of one another then it is not possible to calculate the amount of information conveyed by any symbol. We propose that these probabilities are estimated by the receiver based on the previous symbols communicated by the transmitter (and vice versa). In this framework, both transmitter and receiver keep a probabilitistic model of the communication process, and update this model as each new symbol is communicated through the channel. Words that have been used recently become more expected, i.e., more "available" in the mind, and therefore less informative.

This would explain why long CNCs appear so often in scientific articles, and also would highlight the predictions of the UIDh about the presence of CNCs in the scientific register. First, the commonality of CNCs in the scientific register should make them more expected (less informative, easier to understand) both for writers and for readers. In other words, we suggest that CNCs convey *less* information in scientific articles than they normally would in other contexts. Since it is not clear how to estimate the amount of information present in a given CNC, we do not investigate this prediction in this paper.

Second, especially for those CNCs that are not easy to understand in isolation and therefore *do* constitute peaks of information density, the context should play an important role in clarifying their meaning. Assuming that the authors' goal is to communicate reliably with the readers, these CNCs should, as suggested by Bhatia (1992), only be used in situations where the context is more strongly supportive of their appearance.[3] This support could be produced in any number of ways. For example, authors could slowly construct a CNC such as *listener network rank* over the course of several paragraphs, by first juxtaposing its constituent words in smaller structures (e.g., speaking of a *listener network*, and of ranks of such networks), until finally reaching the full CNC form.

Third, as also suggested in Bhatia (1992), CNCs should appear more often toward the end of the texts. This follows from Genzel and Charniak (2002). If the amount of information of a given word w_i is dependent on its context, then we can break the context into two pieces: The *local* context, containing the information of the present sentence; and the *global* context, containing all other sentences. Assume that the amount of information $I(w_i)$ transmitted by each word w_i is constant. As the global context increases, more and more words become predictable based on it. In order for $I(w_i)$ to remain constant, the information contained in each word *locally*, disregarding the global context, needs to increase too, to compensate for how predictable these words have become when we do consider the global context.

---

3 It may seem circular to assume that harder-to-understand CNC "*do* constitute peaks of information density". This is not the case. In fact, the reasoning would only be circular if we assumed that these CNCs exceed the channel capacity (and that their difficulty arises from this fact). We do not make that assumption.

Intuitively, when considered in isolation, we argue that the probability of harder-to-understand CNCs is at least on average lower than the probability of easier-to-understand CNCs, and therefore the first do convey more information than the latter, irrespective of what the UIDh has to say about it. We do not know for sure if these CNCs do "exceed the channel capacity", but this is not necessary for the predictions in the text to follow from the UIDh: if speakers do convey information at a constant rate, then the context should still be more helpful in clarifying the meaning of harder-to-understand CNCs than that of CNCs that are easy to understand.

This has been shown to hold for written text by Genzel and Charniak (2002). It has also been shown to hold inside paragraphs (Genzel and Charniak 2003), even when sentence length is controlled for (Keller 2004). More recently, this has also been shown to hold in dialogues between two speakers by Xu and Reitter (2018), although they also showed that the amount of local information tends to decrease during topic shifts, presumably because the context becomes less informative in these situations (see also Qian and Jaeger 2011 for a similar finding for written text). Of course, if "language production at *all levels* of linguistic representation is organized to be communicatively efficient", then this should also be true for CNCs.

Finally, as discussed above, once a CNC is used for the first time, we expect readers to update their probabilitistic model of the text they are reading, making the CNCs more "available" (i.e., more expected, less informative) for future reuse. As such, they should become part of the reader's "vocabulary", not requiring much reintroduction, and presumably reappearing often in the text that follows.

1.2 The present study[4]

In this study, we investigate the use of CNCs in scientific articles. We use the UID hypothesis as a basis from which we explore the distributional properties of CNCs. Previous research has shown that the number of CNCs increases with the technical level of the scientific publication: the higher the level, the higher the frequency and complexity of CNCs (Horsella and Pérez 1991). Despite this and other previous studies reporting on the frequency (Biber and Gray 2011) and on the difficulty (e.g., Geer et al. 1972) of CNCs, little is still known about their distributional properties in the scientific register. In particular, it is not clear how the preceding text supports their introduction, how often they are reused, or whether they cluster in certain regions of the article. As discussed above, the UIDh makes clear predictions about these questions. To answer them, we constructed a corpus of scientific articles from high impact journals in different fields, identified their CNCs, and performed a qualitative and a quantitative analysis of the identified CNCs. In our quantitative analysis, we counted the number of CNCs in the different parts of the corpus articles, counted how often they were reused, and how often two-word subparts appeared in the text passage preceding the CNCs (e.g., for the CNC *water treatment facility*, we counted the bigrams *water treatment* and *treatment facility*). In our qualitative analysis, we took a closer look at how CNCs are set up by their context, identifying the strategies used by authors when introducing a new CNC.

4 The data and code used for the analyses reported here are available in OSF: https://osf.io/4a9y5/.

The structure of this paper is as follows. In the next section, we discuss how the corpus data was collected and processed. We then proceed with the quantitative and the qualitative analyses. Finally, we discuss how our results relate to the UIDh and to other theories of sentence processing, and consider how our quantitative measures could be improved and what they reveal about the relation between the UID hypothesis and the use of NCs.

2 Corpus construction

We formed a dataset containing research articles from nine high-impact journals in the fields of Biology, Economics and Linguistics, published either in 2016 or 2017. The choice of these fields was arbitrary, but we purposefully included texts from both the Natural Sciences (Biology), and from the Social Sciences (Linguistics and Economics). For each field, we collected exactly 54 articles, but the number of articles from each journal varied (see Table 1). The full list of articles can be found in the supplementary materials.

We downloaded each article in PDF format and converted it to TXT using the AntFileConverter (Anthony 2017), which outputs text files in Unicode format. We then manually removed all abstracts, headers, references, appendices, tables, notes and pages numbers from the resulting text files. In addition, we manually replaced a number of mathematical formulas from the text files with a more natural textual continuation. For example, the sentence "*To test this idea, we estimate the following regression: FORMULA where all variables have been defined previously*" was changed

Table 1: The list of all journals from which the corpus articles were collected. Impact factors for all journals came from their respective webpages as retrieved in April 13th 2021.

Field	Journal	Impact	Year	Number of articles
Economics	Ecological Economics	4.482	2017	11
	Energy Economics	5.203	2017	20
	Journal of Accounting and Economics	3.723	2016 and 2017	23
Biology	Behavioural Brain Research	2.977	2017	23
	Biological Psychology	2.763	2017	18
	Current Biology	9.601	2017	13
Linguistics	Applied Psycholinguistics Journal of Child	1.412	2017	16
	Language	1.620	2017	24
	Journal of Sociolinguistics	1.630	2017	14

to "*To test this idea, we estimate the following regression, where all variables have been defined previously*".[5] This latter manual text editing step was performed in an attempt to improve the output of the part of speech (POS) tagger, and did not affect results of the further analysis, because the changes were made on function words or punctuations.

We then manually inserted *section tags* in each of the files. In particular, we added the tags <*Introduction*>, <*Middle*> and <*Conclusion*> to the files as follows:

- **Introduction**: added at the very beginning of each file. An *Introduction* extended from the beginning of the file until just before the Methods section. Since articles differ widely in their structure, this may include theoretical sections and descriptions of the related literature.
- **Middle**: added just before the header corresponding to the Methods section. The *Middle* part generally contained the Methods and the Results sections. In articles reporting more than one experiment, the Middle part also included the Discussion sections associated with each single experiment.
- **Conclusion**: added just before the header corresponding to the Discussion section of a given file. In articles reporting a single experiment, the "Conclusion" corresponded to the Discussion and Conclusion sections. In articles reporting several experiments, it contained the General Discussion and the Conclusion sections.

In the rest of this paper, we use the word *section* to refer to each of these article parts.

The TXT files also contained a number of spurious Unicode characters that needed to be cleaned before analysis. Hence, a Python (van Rossum and Drake 2009) script was used to remove all non-printable characters from the dataset, and replace all typographic ligatures with their constituent characters (e.g., the single character representing "ffi" was replaced by two letters "f" and one "i"). This did not remove special characters such as "é" or "ã", which only appeared casually throughout the dataset.

The dataset was then tokenized, part-of-speech (POS) tagged and parsed using the *spaCy* library (Honnibal et al. 2019). The final output of this procedure was a table where each row contained a token,[6] its lemma, its POS tag, its head, the relation between the word and its head, the file where the word came from, the field (Biology, Economics or Linguistics) of that file, the word's position in the sentence, its position

5 In most cases, we just removed the formula, as in the example. Sometimes, we also removed related punctuation (e.g., the sentence *"we use the following model: FORMULA"* was replaced with *"we use the following model."*), and sometimes we also added a pronoun (e.g., the sentence "*Specifically, we estimate: FORMULA where …*" was modified into "*Specifically, we estimate it where …*").

6 We use *word* and *token* here interchangeably to refer to the any form output by the tokenizer including punctuations, hyphens, brackets, etc., and even bound morphemes such as *'s* or *n't*.

in the section, its position in the file, and a unique ID for the whole word in the dataset. This yielded a dataset containing 336,466 words originating from Biology papers, 510,685 words from Economics papers, and 550,992 words from Linguistics papers. For the sake of simplicity and ease of future reference, we call this the Sciper (SCIentific paPER) corpus.

3 Quantitative study of CNCs in research articles

In order to analyse the distributional properties of the CNCs in the Sciper corpus, we needed to first identify them. We start this section by describing the procedure used to identify CNCs. We then follow with an analysis of each of the aforementioned questions: We analyse the position of CNCs across the articles, the number of times CNCs are preceded by their subparts, and the number of times they repeat.

3.1 Identifying CNCs in Sciper

Recall that the corpus is a table containing the words of 162 articles such that each row represents a word. We used this table to count the number of CNCs by article and, inside the articles, by section. Since article length varied widely (see Figure 1) we also calculated the proportion of words pertaining to CNCs for each 1,000 words in the articles.

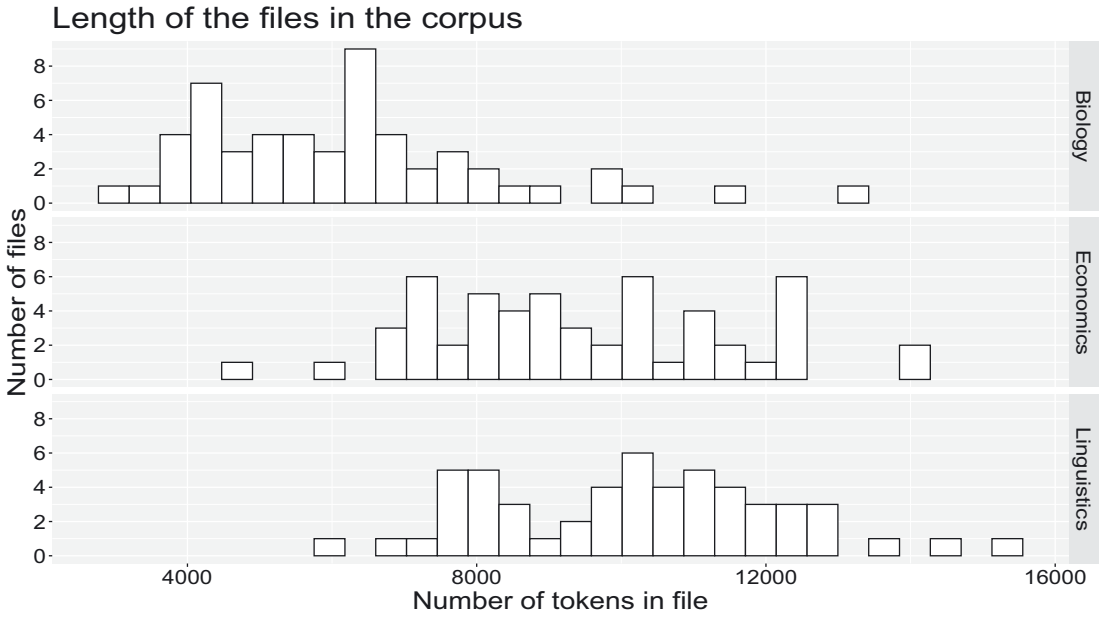


Figure 1: Histogram of the length of the corpus files.

A sequence of words was deemed a CNC if it followed either the structure *(Adjective)+ (Noun){2,}* (at least one adjective followed by at least 2 nouns) or the structure *Noun{3,}* (three or more nouns in sequence).[7] This yielded 24,519 sequences of 3 or more tokens. These sequences found were quite noisy. Many contained incorrectly POS-tagged tokens, tokens with spurious characters (letters in other languages, such as Greek, Chinese or Hebrew), or words from other languages that used the Latin alphabet (such as German, Finnish or French). Sometimes, the construction "ProperName et al." was POS-tagged in a way that would yield a CNC. Sometimes, the tokens were single letter characters (such as "M E T H O D"), or the sequences contained words such as "A." or "i.". Sometimes, the PDF to TXT conversion led to incorrectly separated words (e.g., "inbothFrenchandZulu" or "Na ture"). Finally, many of the articles used cryptic acronyms (like "GABA", "AgCl" or "trkA") or unity values (like "ms" or "kg") that were not of interest to us.

Therefore, we applied the following cleaning procedure to the data. First, we automatically identified and discarded all CNCs containing any non-letter symbols except for the "." (period) character. Examples of non-letter symbols were numbers, Greek letters (mostly used in formulas), hyphens and mathematical operators. If a number appeared preceded by a "." at the very end of the last CNC token (e.g., "earnings response coefficient.9"), we assumed that it corresponded to a footnote, and the CNC was not discarded. In addition, capitalized words containing up to three letters were also removed, as well as words containing 2 characters ended by a ".", or words composed of a single character.

In a final step, all items were manually inspected, producing a number of exceptions to the aforementioned rules (such as "AI", "IQ", "CEO", "US", "GDP" or "EFL"), and a number of words we considered unlikely to constitute CNCs despite conforming to the rules above (e.g., "vs.", "Inc.", "Ltd" or "Rev."). We also manually listed many foreign words (from Dutch, German, Portuguese, French, Finnish, Japanese and other languages) that were found among the sequences, and whenever a CNC contained any of them it was discarded. Finally, this manual inspection also allowed us to find a number of items that conformed to all criteria but were not CNCs. This was typically the case when the item words were, for example, at the border of an adverbial clause (1) or constituted an apposition (2).

(1) *Table 2 shows that animacy significantly affected children's performance: across all* ***sentence types children*** *performed better on sentences that contained inanimate head nouns.* (Kirjavainen et al. 2017: 133)

7 Although we use regular expressions to describe the expected structure of the CNCs, the fact that the data was organized into a table meant that our implementation actually did not use regular expressions to look for CNCs.

(2) *We analyzed the data using mixed-effects logistic regression models to test the effect of four predictors on the continuation selected by participants: sentence type (cleft or canonical), grammatical function of the focus (subject or object), language group (English control, French control, or L2 French) and proficiency for the French data (native controls, low, intermediate, or advanced). Prior to analysis, the two* ***predictors sentence type*** *and grammatical function were effect coded (i.e., sum-coding with values -1 for cleft and +1 for canonical/-1 for object and +1 for subject).* (Destruel and Donaldson 2017: 715)

This cleaning procedure led to the exclusion of 7,599 sequences. The remaining 16,920 CNCs were analysed as reported in the following sections.

3.2 How are CNCs distributed through scientific articles?

To analyse whether CNCs are more common toward the end than toward the beginning of scientific articles, we counted the number and length of CNCs in each section and paper. We analysed these counts using a Linear Mixed Effects Models (LMEM). Following Biber and Gray (2011), we calculated the sum of the lengths of the CNCs (i.e., the number of words of the CNCs) in a given article section divided by the section length and multiplied by 1,000. We refer to this measure as the CNC Proportion. For example, if a section contained exactly 1,000 words, 50 3-word CNCs, and 10 4-word CNCs, then:

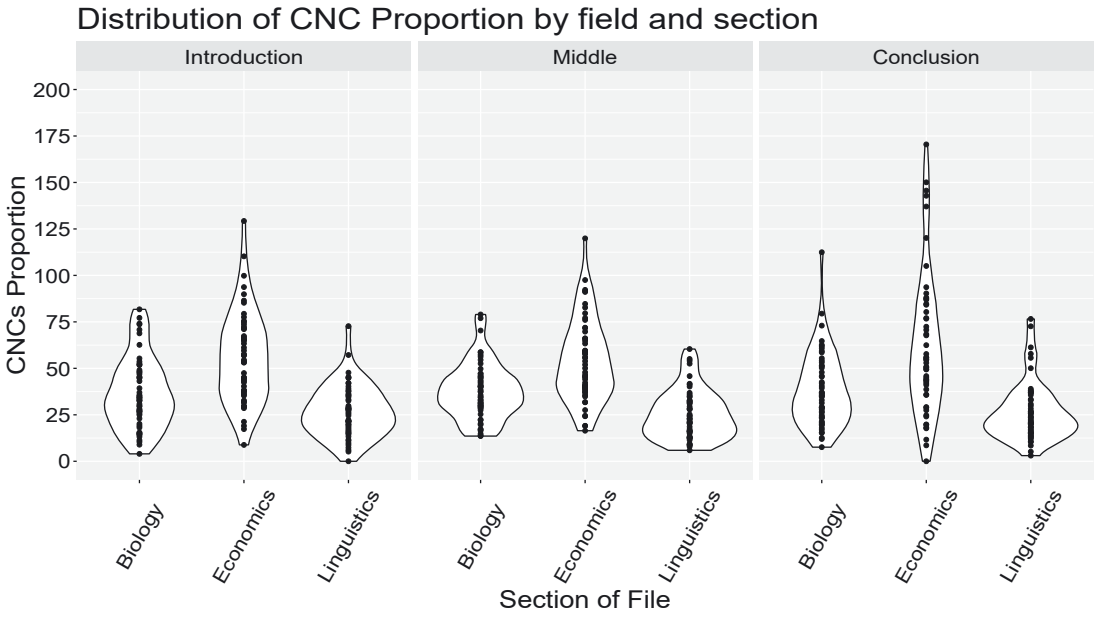
$$\text{CNC proportion} = \frac{(50\,\text{CNCs} \times 3\,\text{words}) + (10\,\text{CNCs} \times 4\,\text{words})}{1000\,\text{words in article}} \times 1000$$

A visual inspection of the distribution of this measure made it clear that it is not normal. Therefore, we log-transformed the data (compare Figure 2a with Figure 2b). This led us to discard two data points (two article sections) because they had no CNCs.

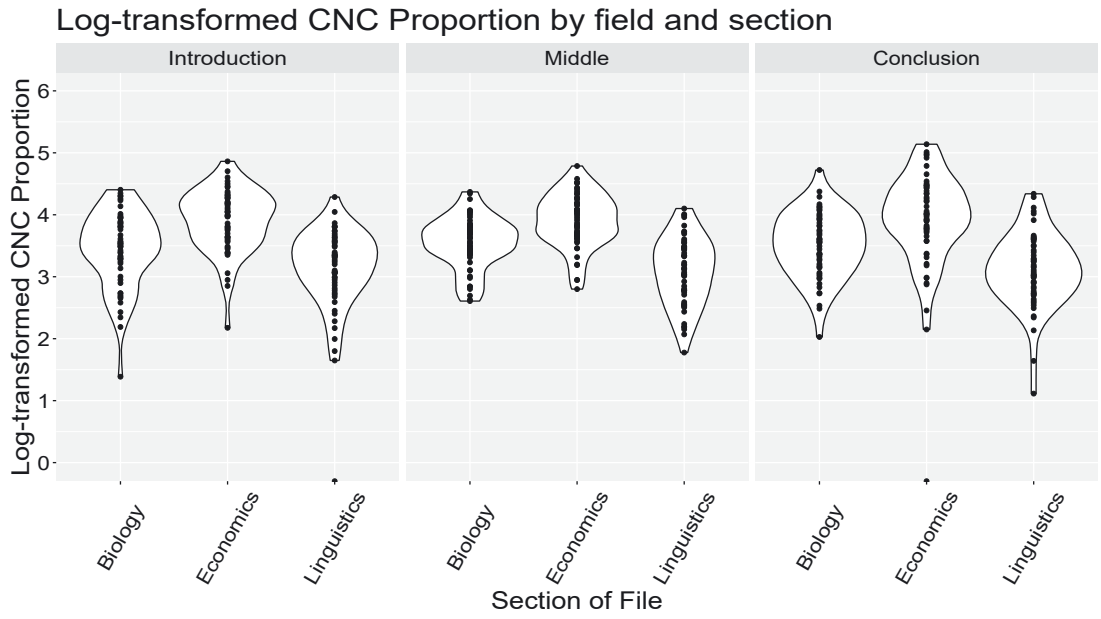
In Figure 2 each data point represents the proportion of CNCs in a given section. Both plots show values calculated for the same data. Linguistics articles seem to have on average a lower proportion of CNCs than Biology, and Economics a higher proportion of CNCs than Biology; but there does not seem to be an effect of the article section.

These results were confirmed in the LMEM whose coefficients are shown in Table 2. The LMEM model was fit using the log-transformed data,[8] with section and

8 We also fit the untransformed data in order to compare the model fit of the two datasets. As expected, the model fit was better with the transformed data.



(a)



(b)

Figure 2: The CNC proportion by field and section. Each datapoint is a file section in the corpus. (a) Raw values; (b) log-transformed values.

field of study as fixed effects, and random intercepts by file.[9] In other words, the model used the formula:

9 Note that it would not be possible to add a random slope by section because the number of random slopes would be the same as the number of files.

Table 2: Results of the linear mixed effects model.

CNC proportion (log-scaled)						
	Estimate	**Std. error**	**df**	**t Value**	**Pr(>|t|)**	
(Intercept)	3.498	0.063	160.573	55.776	0.000	***
Intro vs. middle	0.013	0.040	321.177	0.328	0.743	
Middle vs. conclusion	0.012	0.040	321.177	0.305	0.761	
Biology vs. economics	0.421	0.089	160.821	4.745	0.000	***
Biology vs. linguistics	-0.387	0.089	160.821	-4.359	0.000	***

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

$$\text{DV} \sim \text{section} + \text{field} + (1 \mid \text{file})$$

The Field factor was treatment coded with Biology as the reference factor. The Section factor was difference coded in the order Introduction, Middle, Conclusion.

When controlling for article length, we found no evidence that CNCs cluster around the Conclusion, or that the number of CNCs increases toward the end of the articles, as previously predicted based on the UIDh. We return to this point in the Discussion.

3.3 Are CNCs preceded by their subparts?

In this analysis, we take a first step in understanding the strategies authors use when introducing CNCs. A deeper discussion will be presented in the qualitative analysis.

For each CNC, we count the number of bigram subparts that appeared in the whole text before the CNC. For example, given the CNC *left feeder reward probability*, we counted how many times *left feeder*, *feeder reward* and *reward probability* appeared in the whole portion of the article preceding it.[10] We count *bigram* subparts because they are the simplest structure we could count that is more complex than a word.

Figure 3a shows the distribution of the CNCs by the number of subparts before them. Since CNCs at the end of a paper have much more text before them, we also calculated the distribution only considering the CNCs in the Conclusion section (all of which have, of course, a large portion of text preceding them; Figure 3b).

10 The script that performed this count did not take into account any other source of information and did not consider the possibility that the CNC may appear as a whole in the preceding text. That is, if the entire CNC *left feeder reward probability* appeared in the preceding text, then the script would consider each of the subparts *left feeder*, *feeder reward* and *reward probability* as a "match", therefore summing 3 to the final count.

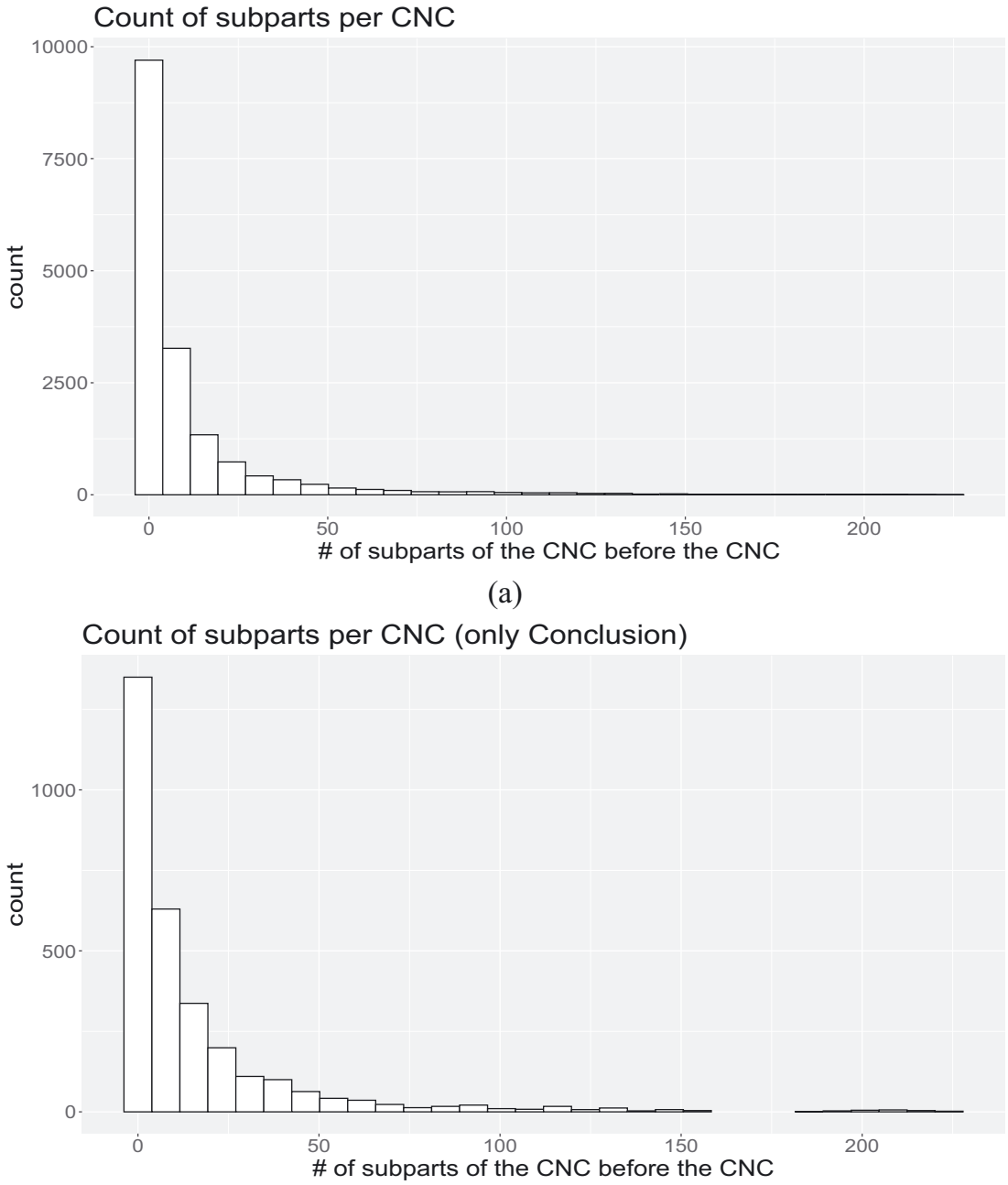


Figure 3: The distribution of two-word subparts of the CNCs preceding the CNC. (a) Histogram of number of subparts found in the whole article text preceding the CNC. (b) Histogram of number of subparts found in the whole article text preceding the CNC (only considering CNCs in the Conclusion).

As can be seen, very few subparts appear before the vast majority of the CNCs in our corpus. A third of the CNCs in the corpus (34.1 %) are not preceded by any subpart at all. This percentage increases to 61.6 % if we add to this set the CNCs preceded by their subparts once (10 %), twice (8.16 %), thrice (5.05 %), four times (4.21 %) and five

times (3.26 %). Analyzing it from the opposite perspective, less than 20 % of the CNCs (19.47 %) are preceded 15 or more times by their subparts.[11] This was unexpected, but we return to this matter in the qualitative analysis below.

3.4 Do CNCs repeat often after their first use?

As discussed above, authors such as Salager (1984) suggested that CNCs are 'ad hoc names', used to refer to new concepts. In addition, we suggested earlier that words that have been used recently become more "available" in the reader's mind. Both views predict that, once a CNC has been used for the first time, it would enter the reader's vocabulary and be reused again and again in the scientific article.

To test this possibility, we counted the number of times each CNC appeared in the same article. Table 3 shows the results of this count. There was a total of 12,136 unique CNCs in all articles of the corpus.[12] Interestingly, the vast majority of CNCs (83.7 %)

Table 3: Number of CNCs by number of occurrences. The same CNC was counted separately if it appeared in a different article.

Number of occurrences in the same article	Number of CNCs (exact matches)	Number of CNCs (after stemming)
1	10,158	9,731
2	1,189	1,184
3	363	382
4	145	158
5	73	92
6	53	59
7	32	26
8	28	28
9	15	19
10	16	11
More than 10	64	75

11 We also tried stemming and decapitalizing each CNC. For example, after this processing, the CNC *Drosophila DRPLA models* would become *drosophila drpla model*. Then we also stemmed/decapitalized all sequences $word_1word_2$ in the corpus and looked for all cases in which they matched. This yielded a very similar distribution.

12 Note that the same CNC was counted twice if it appeared in two different articles. For example, the CNC *oil price volatility* appears in two Economics articles, and was therefore counted two times. The rationale for this was that the context of the CNCs (i.e., the information that presumably sets up the CNC so that it can be succesfully interpreted) would not be shared between the two articles.

Table 4: Distribution of CNC Occurrences according to CNC length.

Number of occurrences in the same article	Number of CNCs with length 3	Number of CNCs with length 4 or more
1	8,211 (82.47 %)	1,947 (89.31 %)
2	1,023 (10.28 %)	166 (7.61 %)
3	325 (3.26 %)	38 (1.74 %)
4	133 (1.34 %)	12 (0.55 %)
5	66 (0.66 %)	7 (0.32 %)
6	49 (0.49 %)	4 (0.18 %)
7	29 (0.29 %)	3 (0.14 %)
8	28 (0.28 %)	0 (0.00 %)
9	15 (0.15 %)	0 (0.00 %)
10	16 (0.16 %)	0 (0.00 %)
More than 10	61 (0.61 %)	3 (0.14 %)

only occurs once in a given article, and this proportion grows to 96.5 % if we add to this set the CNCs that appeared twice (9.8 %) and three times (2.99 %) in the same article. Table 3 also shows how the results change if we apply stemming/decapitalization to the CNCs before performing the count. Based on this data, CNCs are *more often than not* "disposable" words.

We also considered the possibility that the length of these structures may have an impact on whether they may become an ad hoc name or not. Arguably, on average, the longer the CNC the "denser" it is: a CNC such as *unconditional mean audited statement collection rate* represents probably a higher peak in information density than something like *Chinese stock market*. Hence, it may be that the distribution of short CNCs would be different from that of longer ones. Table 4 shows the distribution of 3-word CNCs compared to that of CNCs composed of 4 or more words. As can be seen, although 3-word CNCs are much more frequent, the distributions of the two groups are quite similar.[13]

In summary, in our quantitative analysis, we found no evidence that CNCs cluster in certain parts of the scientific articles. In addition, CNCs are not often preceded by their bigram subparts, and are not often repeated after their first use. We discuss these results in more details in the General Discussion below.

13 Similar results were found for the subpart count.

4 Qualitative analysis of CNC use

We then analysed CNC use from a qualitative perspective. We investigated the way in which CNCs are preset by their context. In particular, we looked for strategies authors might employ to introduce the CNCs in a way that makes their understanding easier. Since only 3.5 % of CNCs occur more than three times in a given paper, we decided to investigate what differentiates these CNCs from the others. Therefore, in the description below, we make a distinction between "recurrent" CNCs (those that appeared four or more times in a paper) and "disposable" CNCs (those that appeared up to three times). We randomly selected 50 disposable and 26 recurrent CNCs from our corpus for manual analysis. This produced a total of 259 CNC occurrences. We selected fewer recurrent CNCs because by definition they led to a much higher number of occurrences, that had to be analysed individually (see Table 6 for a list of all CNCs analysed, as well as the number of times they appeared in a given article). For each CNC occurrence, we examined the text surrounding the CNC, looking for similarities between the different CNC uses in our dataset. Three items (all of which were disposable CNCs) had to be discarded: Two of them were not CNCs upon closer analysis, and one of them was in a part of the corpus in which the PDF to TXT conversion did not work properly.

From this dataset, a number of strategies emerged that authors seem to commonly employ when introducing CNCs. Although some CNCs may not be perfectly categorized in one or the other strategy, we believe this is still a useful first step toward explaining the data. Table 5 shows the number of CNC occurrences that were categorized according to each strategy. As can be seen, the distribution of strategies was substantially similar for both recurrent and disposable CNCs, both in their first occurrence and whenever they were reused. In other words, we found no factors that explain what specifically distinguishes the CNCs that occur often in the articles from those that appear only a few times. In the following, we describe each of the strategies.

4.1 CNC first use

4.1.1 Gradual presetting

The vast majority of the CNC use in our dataset (see Table 5) followed a *gradual presetting* strategy, in which the authors start by using simpler constructions or introducing the meaning of certain words, which are then slowly put together into more complex structures. When the reader arrives at the CNC, the CNC is already

Table 5: The number (and proportions) of CNCs categorized according to each strategy.

CNC introduction – first use		
Strategy	**Recurrent CNCs**	**Disposable CNCs**
Gradual presetting	17 (65.38 %)	30 (63.83 %)
No preset: not required	1 (3.85 %)	4 (8.51 %)
No preset: general scientific lingo	0 (0 %)	3 (6.38 %)
No preset: field terminology	7 (26.92 %)	8 (17.02 %)
No preset: refer to paper	1 (3.85 %)	2 (4.26 %)
Total	26	47
CNC reuse		
Simple reuse	119 (70.00 %)	10 (76.92 %)
Long-distance reuse	51 (30.00 %)	3 (23.08 %)
Total	170	13

established as a natural shorthand for the structures that preceded it. As can be seen in (3), this kind of CNC presetting happens over the course of several paragraphs, some of which discuss topics that are only tangentially related to the final meaning of the CNC.

(3) *Although newborns begin life with the ability to discriminate both native and non-native phonological contrasts attested in the world's languages* (…), *their ability to discriminate non-native consonants and vowels gradually declines between 6 and 12 months* […]
Such phonological contrasts of different spoken languages can be divided into segmental units, including consonants and vowels, and suprasegmental units, such as stresses and ***tones****. Almost 60–70% of the world's languages are* ***tone*** *languages,* […] *Moreover, developmental changes in* ***tone perception*** *were systematically explored by Yeung, Chen and Werker (2013). Their results demonstrated that language experience might affect the* ***perception*** *of lexical* ***tones*** *as early as 4 months: English-, Cantonese-, and* ***Mandarin****-exposed infants each demonstrated different discrimination abilities that accorded with the properties of their native language at this stage.* […]
Modern ***Mandarin*** *is a* ***tone*** *language with relatively simple syllable structure* […]. *The phonological saliency hypothesis (Hua and Dodd 2000) might account for the order of phonological production in* ***Mandarin****, with* ***Mandarin tones*** *being the most salient.* […]
For ***Mandarin tone perception****, although the study of …* (Chen et al. 2017: 1414–1416)

Table 6: A list of all the CNCs analysed and the number of times they occurred in a given article.

Recurrent CNCs		Disposable CNCs			
Academic language comprehension	4	Analyst coverage proxy	1	Nearest neighbor distance	1
Adverse selection costs	11	Auditory cue use	1	New York City	1
Audit quality metrics	6	Average familiarity score	1	Peking University Health Science Center	1
Chinese stock market	16	Baseline comparison condition	1	Picture-word naming times	2
Chronic restraint stress	6	Binyan verb pattern	2	Posterior scalp sites	1
Emotional stop signal task	7	Climate change policy	2	Potential benchmark sets	2
Energy tax reform	25	Different accouting standards	1	Previous business relation	1
Final pitch contour	5	Drosophila DRPLA models	2	Reliable stop signal reaction time	1
Financial reporting quality	4	Effective communication tool	1	Short run impacts	3
Firm-specific cash flows	6	Energy generation data	2	Significant standard deviation parameters	1
Higher audit quality	5	Eye scanning pattern	1	Single lever press	1
Initial PCAOB inspections	4	False belief test trials	3	Slang-heavy speech style	1
Mandarin tone perception	4	Functional connectivity study	1	Specific brain areas	1
Novel label trials	4	Gene sequence divergence	1	Specific comprehension subskills	1
Oil return volatility	17	Heartbeat perception task	2	Statement collection rate	1
Positive policy surprises	5	High frequency asymmetries	1	Stronger consumer reponse	1
Public transport interchange	4	Higher cognitive control demand	1	Successful task completion	1
Retrosplenial cortex lesions	11	Higher completion rates	1	Theoretical disclosure literature	1
Risk factor disclosure	9	Human research ethics committee	1	Trochaic target words	1
Same auditor practice office	6	Implicit gender influences	1	Voltammetric recording site	2
Securities offering reform	4	Linear programming algorithm	1	Western sexual identity terms	1
Syntactic frame diversity	10	Lowest energy usage	1		
Upcoming policy news	4	Main clause subject	1	**Discarded**	
Vehicle fuel economy	7	Mammalian suprachiasmatic nucleus	1	Education advanced level	1
Violent offender sample	4	Marginal water consumption	2	Index character types	1
Word learning task	8	Natural spring habitats	1	Simon task performance	1

It is important to note that "Gradual Presetting" does not necessarily mean that the CNC was easily understandable at the point of its use. Many CNCs in the Gradual Presetting category were quite "jargon heavy", as example 4 shows.

(4) *Polyglutamine (polyQ) diseases are neurological conditions due to an expanded CAG repeat resulting in polyQ stretches in the encoded protein. This family of disorders includes Huntington's disease, dentatorubral-pallidoluysian atrophy (**DRPLA**), and several spinocerebellar ataxias. **DRPLA** is caused by the expansion of a CAG stretch in the ATROPHIN-1 (ATN1) gene* […].
*Several **DRPLA** mouse **models** have been previously generated, all recapitulating important aspects of the disease [11–13]. We have predicted dysfunctional autophagy from previous **Drosophila** studies on **DRPLA*** […the text goes on about DRPLA mice for two paragraphs …]
*Previous **Drosophila** studies indicated a blockage of autophagic clearance in **DRPLA*** […six more paragraphs on other things …]
*Despite robust similarities indicating block at lysosomal level as observed in **Drosophila DRPLA models**, in **DRPLA** mice we detected additional events that …* (Baron et al. 2017: 3626–3630)

Examples (3) and (4) also illustrate how counting the number of bigram subparts in the whole text preceding the CNC may not be a good way to investigate the way CNCs are introduced: In the more than two pages of text preceding the *Mandarin tone perception* CNC, and although the text clearly establishes its meaning, there are only two occurrences of *tone perception*, and five occurrence of *Mandarin tones*.[14] Even more dramatically, the two text pages preceding *Drosophila DRPLA models* have not a single occurrence of *DRPLA models* or *Drosophila DRPLA* (it does have two occurrences of *DRPLA mouse models*, which would not be considered with the bigram counting strategy).

In addition, note that it is not only the words composing a given CNC that are relevant for its presetting. In some cases, other words, semantically related to the CNC words, may have an influence on our expectations. In (4), the word *studies* is used in a similar sense to the word *models*, indicating that both can be combined in similar ways to form complex structures. Similarly, since it is clear that the *Drosophila* is an animal, words such as *mouse* and *mice* may also be used as analogies

14 Note that this is a plural, and would only be correctly counted when the stemming operation was applied prior to the counting procedure. Without this stemming operation, none of the (five) occurrences would have been considered.

in potential combinations with *Drosophila*. We return to this topic in the General Discussion.

4.1.2 No presetting

As Table 5 shows, we found a number of CNCs in our dataset that were used without any previous explanation about its words. In these cases, authors varied substantially in how much they seemed to expect readers to know about the CNC meaning. We further divide these strategies into four subcategories: *No presetting required*, *general scientific lingo*, *field terminology*, and *refer to paper*. We discuss each of these strategies below.

4.1.2.1 No presetting: not required

In some cases, the CNCs were very easy to understand despite very little surrounding contextual information. These are composed of familiar words organized in familiar structures. In some cases, contextual information is necessary, but only to disambiguate the various possible senses in which a given word can be used. For example, the word *energy* may assume several senses in a CNC such as *lowest energy usage*: in a Biology context, a cell may require "energy" to live, and there may be a type of cell that has the "lowest energy usage" in the body; in a game, a character may have skills that can only be used when it has enough "energy", and there may be a skill that requires the "lowest energy usage"; in an Economics sense, "energy" could be interpreted as "what is needed for things to move", and a type of car may have the "lowest energy usage" among all possible cars.

The four disposable CNCs that were categorized in this group were *specific brain areas*, *Peking University Health Science Center*, *average familiarity score*,[15] and *lowest energy usage*.[16] The only recurrent CNC was *Chinese stock market*.

4.1.2.2 No presetting: general scientific lingo

In these cases, the CNC was generally composed only of words that are commonplace in scientific texts, related, for example, to the study design, to statistics or the

15 This was a Biology article (Goto et al. 2017) in which participants rated the worth of certain products (how pleasant they were, how much they wanted them, how familiar they were with the product, and whether they would buy them or not) while having their EEG waves recorded. At the point where the CNC is used, the word *familiarity* was not explained. The explanation is only given at the end of the page, in a separate subsection, when the task the participants performed is described.
16 This is in a context where the article is speaking about the energy used by several countries, among which "Morocco has the lowest energy usage".

previous work done on the topic. The three CNCs pertaining to this category were *baseline comparison condition*, *theoretical disclosure literature*[17] and *significant standard deviation parameters*.

4.1.2.3 No presetting: field terminology

Sometimes, the CNC was not introduced, but its meaning was clearly assumed to be understood based on the reader's knowledge of the field. Of course, from the point of view of the writers, it is possible that there is no difference between this and the previous substrategies: the writer would assume some amount of knowledge from the reader, and write based on this assumed knowledge. However, we believe that the type of knowledge is different: in one case, the knowledge is broad and accessible to novice readers; in the other, it is field specific and highly technical. Of course, what counts as "field terminology" is open for debate. For example, should the CNC "linear mixed-effects model" be viewed as just "general scientific lingo" (since it is used in many different scientific fields), or should it be considered "field terminology" when it is used in a Psycholinguistics article? Here, we categorized as "field terminology" anything that would likely not be obviously understood by readers of the other fields in our corpus. Example (5) shows a dramatic example from Biology where the CNC appears in the second paragraph of the article (i.e., the example contains the whole context preceding the CNC).

(5) *Sleep is an essential and evolutionarily conserved behavior from worms to humans [1,2]. It is tightly governed by two independent processes: the circadian clock that determines the timing of sleep and the homeostatic mechanism that controls the amount and depth of sleep [1, 3]. The circadian clock contains a negative transcriptional feedback loop to synchronize the physiology and behavior of most animals to daily environmental oscillations [4–6]. The timing of sleep can be thought of as an output of the circadian clock. Several molecules, such as melatonin, prokineticin 2, and WAKE, have been identified as clock output molecules that regulate the timing of sleep [7–9].*

17 This is one CNC for which the categories we defined do not work perfectly. The word "disclosure" appears several times in the preceding text, and probably has a well understood meaning in the Economics field, since its meaning is not explained in the article. This would have been an argument for inserting this CNC into the Gradual Presetting category. However, the two words "theoretical" and "literature" were not at all introduced in the preceding text. They are only easily understood because the typical Science reader is used to finding these words in article introductions.

> *The circadian clock also regulates the electrical activity of pacemaker neurons, which modulate the status of sleep and wakefulness [10–13]. In vertebrates and invertebrates, the circadian clock drives antiphase oscillations of sodium and potassium conductance to control the daily cycling of membrane potential in pacemaker neurons [14]. It also drives rhythmic transcription of several ion channels in the* ***mammalian suprachiasmatic nucleus****, including L- and T-type* Ca^{2+} *channels, BK channels, and K2P* K^{+} *channels [15,16].* […] (Li et al. 2017: 3616)

Other examples of field terminology CNCs were *adverse selection costs* (Economics), *final pitch contour* (Linguistics), *chronic restraint stress* (Biology).

4.1.2.4 No presetting: refer to paper

Finally, three CNCs presented no presetting at all, but were followed by a reference, pointing the reader to a source where they could find more information (one of them was recurrent: *emotional stop signal task*; and the other two were disposable: *increased nearest neighbor distance* and *reliable stop signal reaction time*). Two of these were found in the same article, which may mean that this is just the result of the authors' style.

4.2 CNC reuse

As discussed earlier, our initial assumption had been that CNCs are 'names' that can be often reused once they have been introduced. In this qualitative analysis, we classified CNC reuse according to how available in the reader's mind the CNC probably is at the point of reuse. We considered "Simple Reuse" the cases in which the CNC has just been used; and "Long-Distance Reuse" the cases in which other (often unrelated) topics were discussed between the previous uses.

4.2.1 Simple Reuse

Sometimes, after a CNC occurrence, the same CNC reappears in the same sentence, the next sentence, or the next paragraph. These immediate reappearances were categorized as "Simple Reuse" (see Example [6]). In these cases, the meaning of the CNC can presumably be easily retrieved and no further introduction is needed for its understanding. In some cases, we also considered "Simple Reuse" when two CNCs were somewhat far apart, but subparts of the CNC appeared often in the intervening text.

(6) *The significant negative association between aggregate earnings and one-month-ahead returns that becomes insignificant when one-month-ahead* ***policy*** *surprises are included in the regression suggests that the market does not fully anticipate the* ***upcoming policy news*** *in aggregate earnings. To shed further light on this results, we examine whether the FOMC announcement-day returns are also predictable. We find a significant negative association between aggregate earnings and one-month-ahead FOMC announcement-day returns when* ***policy*** *surprise is negative. This association is muted when we control for the one-day* ***policy*** *surprises. This finding confirms that the market does not fully anticipate the* ***policy news*** *in aggregate earnings prior to FOMC announcement. More importantly, it provides direct evidence that the negative aggregate earnings-returns association is driven at least in part by the market's reaction to the* ***upcoming policy news***. (Gallo et al. 2016: 104–105)

4.2.2 Long-Distance Reuse

In other cases, after the CNC was used a few times, it vanished for a long stretch of the text. In these cases, it was common for the text to go into a different direction, discussing other topics not related to the CNC. For example, the CNC may have been used in the Introduction of an Economics article, and then disregarded during the Methods section, where mathematical formulas are introduced along with the data where they are applied. Finally, at the end of the Methods section, the authors may bring back why they are using those formulas/data, reintroducing the CNC. This is precisely the case with the CNC in (6): after appearing twice in the article's introduction (the quoted paragraph is on pages 104 and 105), the CNC (and, in fact, even the word *upcoming*) is not used again until page 113, where the article has already moved on to discuss its results (see Example [7]).

In these kinds of situations, we categorized the CNC reuse as a "Long-Distance Reuse". While the CNC *upcoming policy news* in (7) is quite well reintroduced (the bigram *policy news* appears eight times in the preceding text in the same page as the CNC), this was often not the case. In other words, despite the fact that we applied a different categorization, there was not much difference in the way the two types of CNCs were reused: in most cases, the authors seemed to consider the CNCs clear.

(7) *The significant negative association between aggregate earnings and one-month-ahead returns suggests that the market does not fully anticipate the* ***upcoming policy news*** *in aggregate earnings.* (Gallo et al. 2016: 117)

In summary, this qualitative analysis of 76 CNCs (26 recurrent; 50 disposable) identified five strategies used by authors when introducing a CNC for the first time,

and two ways in which CNCs can be reused. We found no difference in the distribution of the strategies used for recurrent and disposable CNCs. These results, as well as those of the quantitative analysis, are discussed below.

5 General discusion

In the analyses above, we investigated a number of distributional properties associated with complex nominal compounds made up of three or more words. In this investigation, we used the Uniform Information Density hypothesis to make predictions concerning these properties. In particular, we predicted that we would find the following results: CNCs would cluster toward the end of scientific articles, CNCs would be supported by the context in which they are used, and CNCs would be reused often once they are introduced. To answer these questions, we performed a quantitative analysis of the 16,920 CNCs present in 162 scientific articles collected from the fields of Biology, Linguistics and Economics, which in turn were divided into Introduction, Middle, and Conclusion sections, and a qualitative analysis of a small subset of them. The CNCs present in these sections were automatically identified and subsequently filtered using both an automatic and a manual method.

For the first prediction, we fit a linear mixed-effects model using as the dependent variable a log-scaled CNC Proportion (to control for section and CNC lengths). We found no differences in CNC Proportion between Introduction, Middle and Conclusion. In other words, we found no evidence in favor of our initial prediction. However, we found significant differences in the CNC use in the three fields of our corpus: Economics articles had a higher CNC Proportion than Biology articles, which in turn had a higher CNC proportion than Linguistics articles. Future research should examine the specific characteristics of CNC use in each of these research fields.

One could suggest that the lack of differences in CNC use between the sections reflects the audience that writers had in mind when preparing their text: authors could try to make the Introduction and Conclusion friendlier to broader audiences, but to concentrate the study's technical aspects in the Methods and Results sections. Figure 2, however, does not support this idea: the proportion of CNCs in the Middle section is not higher than in the other sections. Indeed, in this case we would still expect more CNCs in the Conclusion than in the Introduction, but that is not what we found.

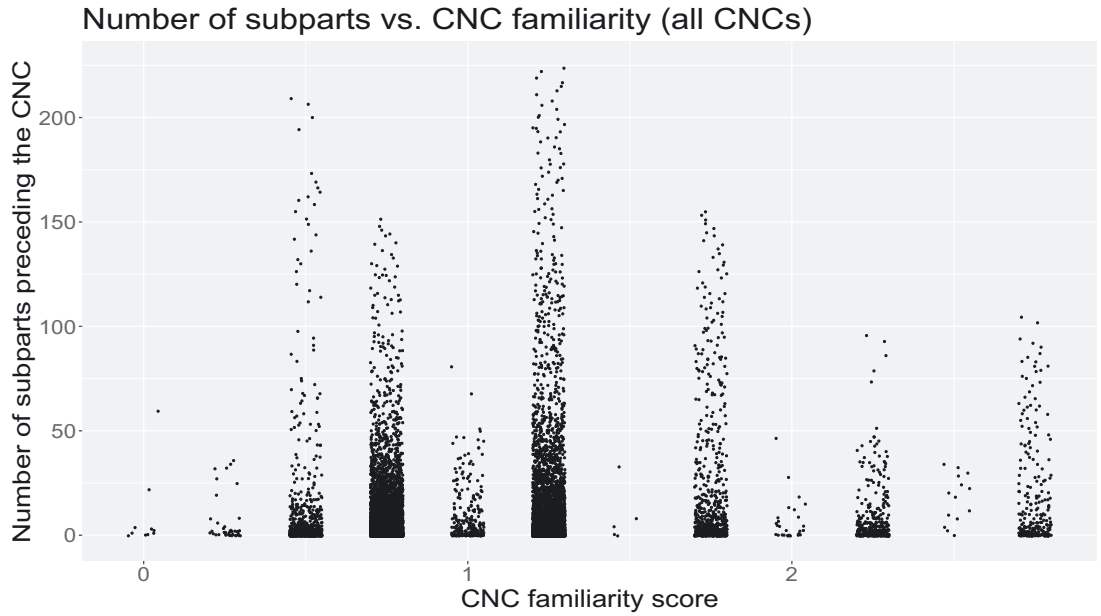
For the second prediction, we performed both a quantitative analysis investigating whether CNCs are preceded by their subparts in the article text, and a qualitative analysis identifying the strategies used by authors to introduce them. From the quantitative analysis, we found that the majority of CNCs are preceded by very

few bigram subparts, and a third of them are not preceded by bigram subparts at all, suggesting that CNCs were not supported by their context.

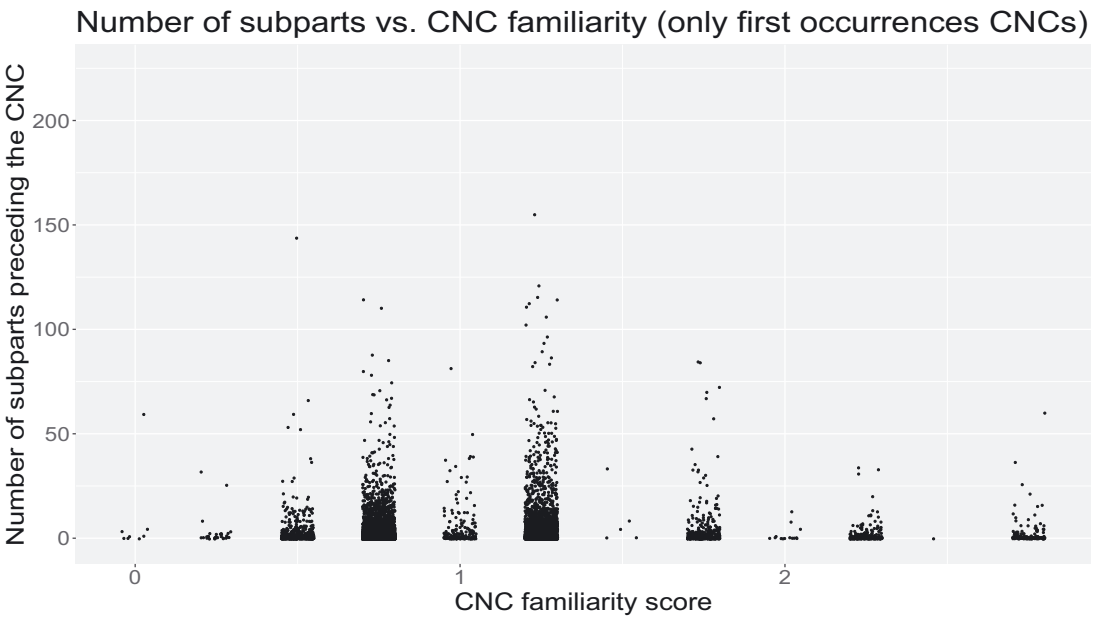
A better picture of the contextual support provided to CNCs was then acquired through the qualitative analysis. It showed that CNCs *are* supported by the context, but in ways that are more sophisticated than simple word-pattern repetitions, which therefore could not be captured by our quantitative measure. Indeed, the majority of the CNCs examined were gradually preset over the course of several paragraphs, with strategies that included embedding in more complex structures single words (rather than bigrams) that were ultimately part of the CNC, or using semantically similar words to hint at the CNC meaning. Among those that were *not* introduced, there was a number of CNCs that constituted either field or scientific jargon, so that (we assume) the author(s) considered that explicit introduction was not necessary. This pattern is consistent with the predictions of the UIDh. Once the simpler structures are introduced, we assume that the mental model kept by the reader about the linguistic characteristics of the text is updated, making them more 'available', increasing the probability that these structures will recur, and therefore reducing their informativeness. With the simpler structures conveying less information, the channel capacity would be underused if newer and more complex structures are not introduced that carry more information.

Note that the scientific articles we analysed were collected from high impact journals. While these may be representative of what 'good' scientific articles look like, they may not be representative of the scientific register as a whole. That is, it is possible that the fact that the CNCs are normally gradually preset in the analysed articles is an artifact of their higher impact, and that scientific articles published in lower impact journals may contain a higher proportion of CNCs used in contexts that are not helpful for their understanding. Indeed, we believe that articles that are judged as 'hard to read' may be 'hard' precisely because they contain frequent peaks of information, many of which may involve CNCs.

Throughout this study, we made assumptions about the difficulty associated with CNCs, claiming that, based on the UIDh, a CNC should only appear in a scientific paper if its context introduces it enough. We made no distinction between CNCs such as "heart rate variability" (quite familiar to a lay reader) and "start arm barrier" (not common at all). Indeed, we had no way to assess the real difficulty associated with any given CNC. Therefore, in the future it may be useful to approach this question experimentally, comparing the difficulty participants experience when confronted with familiar versus unfamiliar CNCs, presented in context versus in isolation. In addition, given the difference in familiarity between the aforementioned examples, it is likely that the first would require much less contextual introduction (and would cause a lower density peak) than the latter. This may have



(a)



(b)

Figure 4: Points represent CNCs with a given familiarity score and a given number of preceding subparts. While in (a) we consider all 3-word CNCs (14,357 data points), in (b) we consider only the first occurrence of a CNC in a given article (9,956 data points).

been one reason why so many CNCs were not preceded by bigram subparts.[18] In order to consider this possibility, we looked up the CNCs of our corpus on Wikipedia, assuming that, if a sequence has its own Wikipedia article, then it is familiar. Denote

18 We thank one anonymous reviewer for highlighting this point.

by $inW(w_{1,\ldots,n})$ a function that is 1 if the sequence of words $w_{1,\ldots,n}$ has its own article in Wikipedia, and 0 otherwise. We calculated a familiarity score f for each CNC composed of words w_1, w_2 and w_3 using the formula below, producing scores ranging from 0 to 2.75, composed by the sum of three terms: one in which the entire CNC is looked up in Wikipedia (which will be either 0 or 1); a second term in which the CNC's bigram subparts are looked up in Wikipedia (resulting in 0, 0.5 or 1, depending on how many bigram subparts have their own pages); and a final term corresponding to whether each of the CNC words have their individual Wikipedia articles (producing 0, 0.25, 0.5 or 0.75).[19]

$$f = inW(w_{1,2,3}) + \frac{inW(w_{1,2}) + inW(w_{2,3})}{2} + \frac{inW(w_1) + inW(w_2) + inW(w_3)}{4}$$

Figure 4 shows how the number of preceding subparts relates to CNC familiarity. In all familiarity levels (except for $f = 2.5$, for which we only have a single CNC repeating 14 times), the median number of preceding subparts is between 1 and 4, i.e., very close to zero. While there is no clear trend indicating that more familiar CNCs are preceded by fewer subparts, we cannot rule out this possibility either. Future studies should consider in more detail the effect of familiarity on the degree to which CNCs are preset. Since familiarity may change over time (a CNC like "database management system", familiar today, barely existed 60 years ago), it might be useful to include date of publication as an additional covariate in such an analysis.[20]

Given the failure of the quantitative measure we used to capture the kind of preset used by authors when introducing CNCs, and given the complex characteristics we found during the qualitative analysis, we believe that a better quantitative measure might have been the number of semantically related words preceding the CNC. Consider (8), where the target CNC is *high reward probability side*, the previous occurrences of the CNC words are bold, and the semantically related words are underscored.[21] In addition, each underscored word is tagged with a number

19 Several CNCs are composed of pluralized words (e.g., "greater information asymmetries").

Wikipedia often redirects these plurals into the singular pages (e.g., "information asymmetries" leads to "information asymmetry"; and "language group" and "language groups" both lead to "language family"), but not always ("gender distinction" leads to "Sex-gender distinction", but "gender distinctions" does not exist). Surprisingly, sometimes pages in the pluralized form do not have a singular version ("Brain regions" leads to the article "List of brain regions in the human brain"; but "Brain region" does not exist). For this analysis we simply search for the words exactly as they are in the corpus, without considering this variability. In addition, we restricted our analysis to 3-word CNCs (a total of 14,357, i.e., 84.85 % of the data) because comparing CNCs of different lengths would be unfair, given the formula we used.

20 We again thank one reviewer for this suggestion.

21 We manually decided which words are semantically related to the compound words. An actual implementation may use, e.g., word vectors produced by an NLP model.

indicating the CNC word that it is related to (for example, the word *low* is tagged with a 1 because it is related to the first word of the CNC, *high*). It is clear that the CNC *high reward probability side* is preset not only by having *high reward*, *reward probability* and *probability side* appear before it, but also by having contextual cues that imply the possibility that these word combinations are likely to occur. For example, the words *choice*, *left* and *right* indicate that there is a *side*, even if the word *side* never appears in the article before the CNC. Similarly, upon reading about the *probability of receiving a reward*, the reader is naturally able to expect the wording *reward probability*.

(8) *An additional cohort of female rats (n = 11) performed a reversal[4] learning task in the same operant conditioning chambers. Animals were trained[2] on the same schedule as cohort 1. Trial initiation was self-paced and began with a nose-poke in the central port. A non-informative tone then prompted the rat to select[3] a sucrose delivery feeder[2]. Correct feeder choices[2] were* ***rewarded*** *with sucrose solution. No* ***reward*** *was given for incorrect choices[2]. In contrast to the Competitive Choice[2] Task, the* ***probability*** *of receiving a* ***reward*** *at a particular feeder[2] was fixed over blocks of 60 trials to either a* ***high*** *or low[1]* ***reward probability****. These* ***reward probabilities*** *reversed[4] at the beginning of each block of trials. For example, the left[4] feeder[2] would be the* ***high reward probability side*** *on trials 1–60, and would then reverse to become the low[1]* ***reward probability side*** *for trials 61–120.* (Wong et al. 2017: 138)

Of course, this measure also has shortcomings. For example, it does not take into account sentence complexity, and does not consider how the words are combined in the context. In addition, counting the number of semantically related words is subjective, requires a large amount of resources, and is beset by problems. For example, how can we best define what constitutes the "preceding text"? Should we consider only the section where the CNC appears, or should we consider the whole article? If the whole article, how can we fairly compare longer versus shorter articles? How can we compare a CNC that appears around the beginning of an article with another one that appears near the end? Some of these problems may be solved with computational approaches that can tag large amounts of data and require fewer resources; however, whether these approaches would lead to results that are similar to those produced by humans is an open question. We intend to explore this issue in future work.

As for the last prediction investigated in this paper, that CNCs would be reused often after their first use, we also found no evidence that this is the case. Indeed, the vast majority of CNCs (83.7 %) were never reused. CNCs do not seem to become 'ad hoc names' for new concepts as suggested by Salager (1984), and instead have a very

local use, being immediately discarded. This held true both for CNCs composed of three words, and for those composed of four or more words.

One should be careful before treating this as evidence against the UIDh for two reasons. First, while we did not investigate this possibility in this paper, it is possible that very dense CNCs become acronyms after their first use. Like CNCs, acronym use has substantially increased in the last decades (e.g., Barnett and Doubleday 2020). Note that this was precisely the case with the two most commonly used CNCs in this very paper (i.e., *complex nominal compound* and *Uniform Information Density hypothesis*). Second, it is possible that repeating CNCs actually conveys too little information, and that, instead of repetition, CNCs would actually undergo deletion of some of their words after their first use. For example, a CNC such as *high reward probability side* could be reused as *high reward side* or even *high side* in contexts where the missing words are obvious. If that is the case, then we should find very different results if we use a relaxed version of the quantitative measure used in this paper. We intend to explore this idea in future work.

The three predictions discussed above were made taking into account the fact that the transmitter does not have full knowledge about the expectations of the receiver, and is therefore not able to calculate perfectly the amount of information of each word from the perspective of the receiver. As discussed earlier, we suggested that each communication partner keeps a probabilistic model of the communication process and updates this model as new words are transmitted through the channel. As the reader progresses through the text, the two models would slowly converge toward similar probabilities for most words. But is this really the case? The results of our second prediction point in that direction. As Table 5 suggests, authors seem to organize their articles so that, in most cases, CNCs are only used when they can be understood. This makes it surprising that we did not find more CNCs toward the end of the articles. Further research is necessary to investigate why this was the case.

In summary, in this paper, we found no evidence that CNCs are more frequent toward the end of scientific articles, nor that CNCs become 'ad hoc names' after their first use in an article. These possibilities, however, could not be completely discarded, and thus our findings constitute no evidence in favor of or against the UIDh. On the other hand, we found that CNCs *are* preset in their context, but that the way in which this presetting occurs is complex, and could not be captured by the strict quantitative measure we used. This latter finding constitutes evidence in favor of the UIDh. In addition, this finding suggests that speakers do keep a probabilistic estimation of the communication process, updating this model as the communication unfolds.

6 Conclusions

In this study, we investigated the distributional properties of complex nominal compounds composed of at least three words. To do so, we performed a quantitative and a qualitative analysis of the CNCs identified in our corpus, the Sciper Corpus. Some of our results constitute evidence in favor of the Uniform Information Density hypothesis. From the qualitative analysis, a number of improvements arose that can be made on the quantitative measures we used. In the future, we intend to investigate how the quantitative results differ if performed considering these improvements.

Acknowledgments: We thank Abigail Hodge and Daria Gvozdeva for helping with the scripts used for the corpus construction, and Rhiannon Stewart for help with formatting the final version of this paper. Part of this research was completed as a Master's thesis by the second author. The research reported here was funded by a doctoral fellowship to the first author from the Center for Cognitive Science at the Technische Universität Kaiserslautern, via support from the Rhineland-Palatinate State Research Initiative. Finally, we thank Titus von der Malsburg and John Beavers for useful comments on a previous version of this paper.

References

Algeo, John & Adele S. Algeo (eds.). 1991. *Fifty years among the new words: A dictionary of neologisms 1941–1991*. Cambridge: Cambridge University Press.

Anthony, Laurence. 2017. *AntFileConverter (Version 1.2.1) [Computer Software]*. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software (accessed 27 February 2024).

Baldwin, Timothy & Takaaki Tanaka. 2004. Translation by machine of complex nominals: Getting it right. In Takaaki Tanaka, Aline Villavicencio, Francis Bond & Anna Korhonen (eds.), *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, 24–31. Stroudburg, PA: Association for Computational Linguistics. https://aclanthology.org/W04-0404 (accessed 27 February 2024).

Barnett, Adrian & Zoe Doubleday. 2020. The growth of acronyms in the scientific literature. *Elife* 9. e60080.

Baron, Olga, Adel Boudi, Catarina Dias, Michael Schilling, Anna Nölle, Gema Vizcay-Barrena, Ivan Rattray, Heinz Jungbluth Wiep Scheper, Roland A. Fleck, Gillian P. Bates & Manolis Fanto. 2017. Stall in canonical autophagy-lysosome pathways prompts nucleophagy-based nuclear breakdown in neurodegeneration. *Current Biology* 27(23). 3626–3642.

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

Bhatia, Vijay K. 1992. Pragmatics of the use of nominals in academic and professional genres. In Lawrence F. Bouton & Yamuna Kachru (eds.), *Pragmatics and language learning* (Monograph series 3), 217–230. Urbana, Illinois, USA: University of Illinois. https://eric.ed.gov/?id=ED395531 (accessed 28 February 2024).

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

Carrió Pastor, María Luisa. 2008. English complex noun phrase interpretation by Spanish learners. *Revista Española de Lingüística Aplicada* 21. 27–44.

Carrió Pastor, María Luisa & Miguel Ángel Candel Mora. 2013. Variation in the translation patterns of English complex noun phrases into Spanish in a specific domain. *Languages in Contrast* 13(1). 28–45.

Chen, Fei, Gang Peng, Nan Yan & Lan Wang. 2017. The development of categorical perception of Mandarin tones in four- to seven-year-old children. *Journal of Child Language* 44(6). 1413–1434.

Destruel, Emilie & Bryan Donaldson. 2017. Second language acquisition of pragmatic inferences: Evidence from the French c'est-cleft. *Applied Psycholinguistics* 38(3). 703–732.

Dressler, Wolfgang U. 2006. Compound types. In Gary Libben & Gonia Jarema (eds.), *The Representation and Processing of Compounds Words*, 23–44. New York: Oxford.

Floridi, Luciano. 2009. Philosophical conceptions of information. In Giovanni Sommaruga (ed.), *Formal Theories of Information* (Lecture Notes in Computer Science 5363), 13–53. Heidelberg: Springer, Berlin.

Frank, Austin F. & T. Florian Jaeger. 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the annual meeting of the Cognitive Science Society*, vol. 30. https://escholarship.org/uc/item/7d08h6j4 (accessed 28 February 2024).

Gagné, Christina L. & 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). 71–87.

Gallo, Lindsey A., Rebecca N. Hann & Congcong Li. 2016. Aggregate earnings surprises, monetary policy, and stock returns. *Journal of Accounting and Economics* 62(1). 103–120.

Geer, Sandra E., Gleitman Henry & Gleitman Lila. 1972. Paraphrasing and remembering compound words. *Journal of Verbal Learning and Verbal Behavior* 11(3). 348–355.

Genzel, Dmitriy & Eugene Charniak. 2002. Entropy rate constancy in text. In Pierre Isabelle, Eugene Charniak & Dekang Lin (eds.), *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 199–206. Stroudsburg, PA: Association for Computational Linguistics.

Genzel, Dmitriy & Eugene Charniak. 2003. Variation of entropy and parse trees of sentences as a function of the sentence number. In Michael Collins & Mark Steedman (eds.), *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 65–72. Stroudsburg, PA: Association for Computational Linguistics.

Goto, Nobuhiko, Faisal Mushtaq, Dexter Shee, Xue Li Lim, Matin Mortazavi, Motoki Watabe & Alexandre Schaefer. 2017. Neural signals of selective attention are modulated by subjective preferences and buying decisions in a virtual shopping task. *Biological Psychology* 128. 11–20.

Hale, John. 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*, 1–8. Stroudsburg, PA: Association for Computational Linguistics.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem & Adriane Boyd. 2019. spaCy (Version 2.1.6) [Computer Software]. https://spacy.io (accessed 28 February 2024).

Horsella, Maria & Fresia Pérez. 1991. Nominal compounds in chemical English literature: Toward an approach to text typology. *English for Specific Purposes* 10(2). 125–138.

Jaeger, T. Florian. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61(1). 23–62.

Keller, Frank. 2004. The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In Dekang Lin & Dekai Wu (eds.), *Proceedings of the 2004 Conference on Empirical*

Methods in Natural Language Processing, 317–324. Stroudsburg, PA: Association for Computational Linguistics. https://aclanthology.org/W04-3241 (accessed 28 February 2024).

Kirjavainen, Minna, Evan Kidd & Elena Lieven. 2017. How do language-specific characteristics affect the acquisition of different relative clause types? Evidence from Finnish. *Journal of Child Language* 44(1). 120–157.

Kvam, Anders Martin. 1990. Three-part noun combinations in English, composition – meaning – stress. *English Studies: A Journal of English Language and Literature* 71(2). 152–161.

Levi, Judith N. 1978. *The syntax and semantics of complex nominals*. New York: Academic Press.

Levy, Roger & T. Florian Jaeger. 2006. Speakers optimize information density through syntactic reduction. In Bernhard Schölkopf, John C. Platt & Thomas Hoffman (eds.), *Proceedings of the 19th International Conference on Neural Information Processing Systems*, 849–856. Cambridge, MA: MIT Press. https://proceedings.neurips.cc/paper/2006/hash/c6a01432c8138d46ba39957a8250e027-Abstract.html (accessed 28 February 2024).

Li, Qian, Li Yi, Xiao Wang, Junxia Qi, Xi Jin, Huawei Tong, Zikai Zhou, Zi Chao Zhang & Junhai Han. 2017. Fbxl4 serves as a clock output molecule that regulates sleep through promotion of rhythmic degradation of the GABAA receptor. *Current Biology* 27(23). 3616–3625.

Libben, Gary. 2006. Why study compound processing? An overview of the issues. In Gary Libben & Gonia Jarema (eds.), *The representation and processing of compounds words*, 1–22. New York: Oxford.

Limaye, Mohan & Richard Pompian. 1991. Brevity versus clarity: The comprehensibility of nominal compounds in business and technical prose. *The Journal of Business Communication* 28(1). 7–21.

Maurits, Luke, Dan Navarro & Perfors Amy. 2010. Why are some word orders more common than others? A uniform information density account. In John D. Lafferty, Christopher K. I. Williams, John Shawe-Taylor, Richard S. Zemel & Aron Culotta (eds.), *Advances in neural information processing systems*, 1585–1593. Red Hook, NY: Curran Associates, Inc. https://proceedings.neurips.cc/paper/2010/hash/0c74b7f78409a4022a2c4c5a5ca3ee19-Abstract.html (accessed 28 February 2024).

Montero, Begoña. 1996. Technical communication: Complex nominals used to express new concepts in scientific English-causes and ambiguity in meaning. *The ESPecialist* 17(1). 57–72.

Qian, Ting & T. Florian Jaeger. 2011. Topic shift in efficient discourse production. In Laura Carlson, Christoph Hoelscher & Thomas F. Shipley (eds.), *Proceedings of the 33rd annual meeting of the Cognitive Science Society*, 3313–3318. Austin, TX: Cognitive Science Society.

Salager, Françoise. 1984. Compound nominal phrases in scientific-technical literature: Proportion and rationale. In A. K. Pugh & Jan M. Ulijn (eds.), *Reading for professional purposes: Studies in native and foreign languages*, 136–145. London: Heinemann.

Schmidtke, Daniel, Kuperman Victor, Christina L. Gagné & Thomas L. Spalding. 2016. Competition between conceptual relations affects compound recognition: The role of entropy. *Psychonomic Bulletin & Review* 23(2). 556–570.

Shannon, Claude Elwood. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27(3). 379–423.

Solso, Robert L. & Joseph F. King. 1976. Frequency and versatility of letters in the English language. *Behavior Research Methods & Instrumentation* 8(3). 283–286.

Spalding, Thomas L., Christina L. Gagné, Mullaly Allison & Ji. Hongbo. 2010. Relation-based interpretation of noun-noun phrases: A new theoretical approach. In Susan Olsen (ed.), *New impulses in word-formation*, 283–315. Hamburg: Buske.

Tobin, Martin J. 2002. Compliance (COMmunicate PLease wIth Less Abbreviations, Noun Clusters, and Exclusiveness). *American Journal of Respiratory and Critical Care Medicine* 166(12). 1534–1536.

van Rossum, Guido & Fred L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Wong, Scott A., Sienna H. Randolph, Victorita E. Ivan & Aaron J. Gruber. 2017. Acute Δ-9-tetrahydrocannabinol administration in female rats attenuates immediate responses following losses but not multi-trial reinforcement learning from wins. *Behavioural Brain Research* 335. 136–144.
Xu, Yang & David Reitter. 2018. Information density converges in dialogue: Towards an information-theoretic model. *Cognition* 170. 147–163.

Supplementary Material: This article contains supplementary material (https://doi.org/10.1515/cllt-2023-0028).

How does the context preceding a long nominal compound influence its comprehension difficulty?

Anonymous authors
Anonymous universities

Abstract

Nominal compounds (NC), i.e., structures such as *stock market anomalies* or *world market oil price*, are very common in the scientific register, but have been argued to be hard to process. In this paper, we focus on the influence of the support provided by the preceding context on the difficulty perceived by readers when comprehending NCs. In two experiments, participants filled out a questionnaire in which they read NCs presented either in their original context or in isolation, and responded how hard to comprehend the NCs were. Items were selected from a corpus of scientific papers comprising the fields of Biology, Linguistics and Economics, and contextual support was measured by calculating the number of words in the preceding text that were semantically similar (as quantified by a large language model) to the words in the NCs. The results of Experiment 1 showed no effect of context. However, in Experiment 2, NC familiarity was additionally manipulated. In Experiment 2, while high familiarity NCs were not judged very different in the different categories, low familiarity NCs did show a substantial difference between weak and strong contextual support. In addition, items presented in context did not differ from items presented in isolation.

As the title suggests, this paper focuses on nominal compounds (NC). We use the term nominal compound to refer to structures that, on the surface, look like sequences of nouns, such as *brain cell*, *stock market anomalies*, or *world market oil price* (see, e.g., Gamboa et al 2024a for a review; and Bauer et al., 2013, Chapter 19 for a more general view on compounds). These are especially common in scientific texts, a register in which their use has surged in the last century (Biber & Gray, 2011). As Biber and Gray note, this surge has been accompanied by an increase in length, so that, by the 1950's, 3-word compounds had already become "relatively common, and even [4-word] sequences [were] not unusual" (p. 238). In addition, it has been accompanied by an increase in complexity – as reflected by the fact that "[t]he set of possible meaning relationships" between the words of the NCs "expand[ed] greatly in the late nineteenth century and throughout the twentieth century, associated with the wider range of premodifying nouns." (p. 239).

In this paper, we focus on these longer, more complex NCs, with 3 or more words, that became more popular through the twentieth century. Bartolic (1978) explains their popularity by referring to a principle of word economy (p. 258), suggesting that "the information is conveyed in a more condensed form which has a greater impact upon the reader" (Bartolic, 1978, p. 260). Alternatively (or in addition), Salager (1983) points out that they allow for the creation of a "new concept for which the language code has no name", a "fixed expressions owing [*sic*] a scientific meaning which the individual constituents do not have" (p. 142; see Montero, 1996 for a deeper discussion on these arguments). Their frequent use, however, is surprising, since they should arguably lead to processing difficulties. In this paper, therefore, we investigate the processing of long NCs, considering the role of one additional aspect that might partially explain why they are used: The preceding context. Although the processing of two-word compounds has been the focus of much research[1] (Gagné & Spalding, 2013; Schmidtke et al., 2021; Benjamin & Schmidtke, 2023), compounds made up of three or more words have received less attention in the processing literature (but see e.g. Kuperman et al., 2008; Sikos et al., 2017; Gamboa et al., 2024a; Cohen & Staub, 2014). In addition, while the influence of the context on the processing of compounds has been investigated for two-word NCs (e.g., Cohen & Staub, 2014), little is known about the way the context influences the processing of longer NCs.

There is a number of reasons why NCs in general, and long NCs in particular, should be complicated to process (for a deeper discussion, see Nakov, 2013). First, NCs can present relational ambiguity (the several "possible meaning relationships" mentioned by Biber and Gray above): The relation linking the words of an NC is left implicit, and it is up to the parser to decide the best interpretation. For example, compare a phrase such as "oil from olives" with "olive oil": while in the first case the relationship between the words is overt and clear, this is not true in the latter. In a typical situation, the parser does manage to easily figure out that "olive oil" refers to an oil *from* olives, and, say, that "baby oil" refers to an oil *for* babies (Gagné & Spalding, 2009). But the other, discarded interpretations are still possible, and, in scenarios where they would make sense (say, a horror movie), the parser still has to realize that *baby oil* could, in principle, also refer to an oil *from* babies (see Cohen & Staub, 2014 for an example of how the context influences this parsing). This kind of ambiguity, associated with the relation between the NC words, increases with longer NCs. Research investigating the psychological reality of these relations has mostly focused on two-word NCs, and not really considered longer structures.

Second, NCs with three or more words can present structural ambiguity: As the number of NC words increases, so does the number of possible structures of the NC. For example, three-word NCs can

[1] That research has been intimately related to the processing of compound words (such as snowball or healthcare).

follow an AB+C structure (*sperm bank donor*), an A+BC structure (*police narcotics control*), or even be unclear on their structure (*science degree course*; Kvam, 1990). This number of possible structures grows rapidly along with NC length, and should arguably make longer NCs harder to interpret.

Third, the NC head is typically located at the end of the NC, complicating the incremental integration of the preceding NC words, especially for longer NCs. For example, upon reading a sentence such as (1), the parser is only able to tell that *changes* is not a verb once it reaches *are*. Related to that, identifying the head of longer NCs has been shown to be challenging (Gleitman & Gleitman, 1970; Geer et al., 1972, Limaye & Pompian, 1991): Participants asked to paraphrase NCs sometimes produce paraphrases that treat the first word as its head (e.g., from our own results reported below, the NC *acquisition negotiation process* was paraphrased as *the acquisition of the process of negotiation*).

(1) *This happens because the **chip production market changes** are driven by competition with foreign companies.*

Given that long NCs are so problematic to process, it may be surprising that they are so common, in particular in scientific texts. Could it be that they are actually more understandable than the research discussed above would lead us to believe? Most of the existing research on long NC processing either disregards the support offered by the broader context for the understanding of the NCs (e.g., Gamboa et al., 2024a, 2025) or acknowledges its importance but does not control for its influence (e.g., Sikos et al., 2017). But if the context surrounding a long NC supports it well enough, then the NC might be less difficult to comprehend than one would expect upon seeing it in isolation. A dramatic example is shown in (2).[2] Without the context, the NC *start arm barrier* would be arguably uninterpretable; but it becomes fairly clear when the whole paragraph is taken into consideration. Therefore, in this paper we explore how the context preceding an NC may explain its use, supporting its meaning and making it easier to comprehend. We then explore a way to operationalize this contextual support, and investigate in two experiments how this support influences the difficulty perceived by readers upon encountering them.

(2) *Each session consisted of six trials and all animals completed one session a day. Each session consisted of three correct left and three correct right trials, presented in a pseudorandom order. Each trial comprised two stages, a 'sample run' followed by a 'test run'. At the beginning of each trial, two sucrose pellets were placed in each food well and a metal obstacle was placed at the choice point of the T-maze, thereby closing one perpendicular path (Fig. 8). On a sample run, the animal was placed in the start area and the aluminium obstacle removed, allowing the rat to run down the start arm. Because of the metal barrier blocking the entrance to one of the cross arms, the rat could only enter the one open section. Once the rat had collected the sucrose pellets from the well at the end of the open section, the rat was returned to the beginning area, where it remained for 10s while the barrier at the choice point was removed and the same arm as previously visited was rebaited. The test run started as the **start arm barrier** was raised, allowing the animal a free choice between the two cross arms of the T-maze.*
(Powell et al., 2017; emphasis added)

[2] For the sake of the argument put forward in the subsequent paragraphs, some occurrences of the words start, arm and barrier were replaced with synonyms in the example two. This did not make the text easier to read. If anything, it should have made it harder to read. The full original paragraph can be read in Figure 1.

Gamboa et al. (2024b), investigating a similar structure[3] in a corpus of scientific papers, noted that these structures are often preset (i.e., supported) by their surrounding context. They analyzed the compounds in two ways. Initially, given a compound (e.g., *world market volatility index*), they counted bigram substrings (*world market*, *market volatility*, *volatility index*) in the whole text preceding the compound, all the way from the beginning of the paper (henceforth referred to as the *bigram-counting measure*). They expected this count to reflect how well preset a given compound was, i.e., to be an easy-to-implement measure of contextual support. However, for most compounds in their corpus, they found very few occurrences of these substrings, presumably indicating that compounds are not very much introduced by their context.

In addition to this measure, they also performed a qualitative analysis of a small set of the compounds in their corpus, finding that a large portion (but not all) of them *are* introduced by their context, but in ways that were not captured by the simple bigram-counting strategy. In other words, this bigram-counting measure did not seem to be a good measure of contextual support. Instead, they suggested that a better measure could have been counting the number of words in the preceding context that were semantically similar to those of the compound (henceforth referred to as the *similar-words measure*). For example, consider the compound *start arm barrier* in (2), which is preceded not only by the words *start*, *arm* and *barrier* themselves, but also by words such as *beginning* (similar to *start*), *obstacle* (similar to *barrier*) or *section* (in the sense used in the example, similar to *arm*). These words presumably hint at the meaning of the NC as a whole, so that the reader, upon arriving at the NC, has a pretty decent expectation of how to interpret it. Thus, NC contexts containing a large number of words that are semantically similar to the NC words (including the NC words themselves) would arguably support the NC more than NC contexts containing fewer similar words.

In this paper, therefore, we have two interconnected goals, one methodological and one theoretical. First, in order to quantify the way in which long NCs are supported by their context, we implement the similar-words measure of Gamboa et al. (2024b), discussing a number of practical considerations necessary for its use. We report the amount of contextual support of the NCs in Gamboa et al.'s corpus (as measured by the similar-words count), showing that there is great variability in the amount of support NCs receive in their immediately preceding contexts. We hope that this measure will be used in future studies trying to quantify contextual support.

Second, in order to understand how the preceding context influences the processing of long NCs, we use the similar-words measure to produce items for experiments that compare the difficulty experienced by readers when comprehending NCs with low and high contextual support, presented in context and isolation. The results of these experiments shed some light on how or why NCs can be so frequent in scientific papers. We predict, based on the discussion so far, that readers should have less difficulty processing well-supported long NCs than less-supported ones. If this is not the case, then it may be useful to look at the difficulty perceived when NCs are presented in isolation, since the context may be leveling the difficulty of the NCs, making them all similarly difficult. If well-supported NCs *are well-supported by their context* because *they require support*, then we should expect them to be more difficult to process in isolation than less-supported ones. This latter prediction would be in line with Information Theory-based models of language comprehension (e.g., Surprisal Theory, Levy, 2008; Uniform Information Density Hypothesis, Jaeger, 2010). In the first experiment, participants filled out a questionnaire where they read three-word and four-word NCs with varying degrees of support presented

[3] The structure they investigated has a broader definition, which includes N-N compounds, as well as other compounds containing preceding adjectives (e.g., financial market volatility index).

either in context or in isolation and answered how hard they perceived the NCs to be. In the second experiment, NCs were controlled using a measure of familiarity, since some of the compounds used in the first experiment were noticeably familiar (e.g., *heart rate variability*, *word order patterns*). We expected familiar compounds to be comprehensible regardless of context, producing an interaction between contextual support and familiarity.

Some considerations on counting semantically similar words

Counting the number of words in the preceding context that are similar to the words in the compound may seem deceivingly simple. What exactly it means to say that two words are *similar*, however, is hard to define. For example, it is easy to say that "happy" and "joyful" are similar. But should *right* and *wrong*, or *love* and *hate* be considered similar words? What about *left* and *right*? Resnik (1999) speaks of similarity as "a special case of semantic relatedness: for example, cars and gasoline would seem to be more closely related than, say, cars and bicycles, but the latter pair are certainly more similar." (p. 95) It would seem that we need to answer these questions before we can actually implement Gamboa et al.'s similar-words measure.

One way to avoid these issues is to resort to today's algorithms of language processing (e.g., Mikolov et al., 2013; Devlin et al., 2018; Vaswani et al., 2017). This is the strategy we follow in this paper. These algorithms are trained on large swathes of texts, and, once the training is over, they are able to convert any word that is input to them into a *word vector* (a list of numbers). Typically, these algorithms are based on the Distributional Hypothesis (see Sahlgren, 2008 for a review), i.e., the idea that a large portion of the meaning of a word can be described by how it is used along with other words. They will produce similar vectors for words that occur in similar contexts, i.e., that are normally surrounded by the same other words. So, if *love* and *hate* usually appear surrounded by the same words (compare common sentences such as "I love ice-cream!" and "I hate ice-cream!"), then they both will end up being assigned similar vectors, indicating that they are similar (even if their meanings are arguably opposite, as in the case of *good* and *bad*, or *left* and *right*).

Once the algorithm outputs a vector for a word, we can then use the vector to perform any sort of calculations. One common way to interpret a vector is as an arrow pointing in a given direction (see Figure 2). To assess the degree of similarity between two vectors (two words), we can then measure the angle between them. Similar words (vectors) point in similar directions, forming a small angle. Conversely, dissimilar words will point in very different directions, forming larger angles. Typically, we use the cosine of the angle to measure similarity, since it ranges from -1 to 1, and the smaller the angle the closer the cosine is to 1.

The problem thus becomes one of deciding when is it that an angle is "small enough" for two words to be considered similar. In this paper, we made this decision empirically, by choosing the angle that best mimicked the annotations performed by humans on a small number of long NCs.

Determining the strength of contextual support for NCs[4]

Having a clear procedure to automatically count the number of semantically similar words in the context of an NC, we then applied the similar-words measure to all the long NCs of the corpus used by Gamboa et al. (2024b), aiming to explore how much support NCs receive from their context. The corpus is a dataset of 162 journal articles in English containing 1.398.559 tokens collected from the fields of Biology

[4] All scripts used in this paper are available in:
`https://osf.io/6p2c3/?view_only=52aedbb4a3c845549a9c8cfeca08c359`

(336.628 tokens), Economics (510.847 tokens) and Linguistics (551.084 tokens), POS-tagged using the Python *spaCy* library (Honnibal et al., 2020). We automatically identified 6006 sequences of three or more words that were tagged in the corpus as nouns, out of which we automatically excluded sequences containing one-letter words (e.g., "generator type g"), "non-English characters" (characters that are not letters, spaces or hyphens; e.g. "+ age rating"), or spaces around hyphens (e.g., "letter co - occurrence"). This led to a final pool of 5234 NCs (see Table 1), whose context was analyzed.

Table 1. *Total number of NCs in the corpus.*

	3-word Compounds	**4-word Compounds**	**Longer Compounds**
Biology	1077	96	9
Economics	2830	238	25
Linguistics	903	51	5
TOTAL	4810	385	39

In order to make the contexts comparable, contexts were restricted to a text snippet of roughly 200 words[5] before the NCs. A compound's snippet was constructed by locating the token positioned 200 tokens before the compound, and then finding the beginning of its sentence (see Figure 1). Every snippet ended at the end of the NC's sentence. Out of the 5234 NCs, 134 appeared too early in the articles to have a preceding context of at least 200 words, and were therefore discarded. This left a total of 5100 that were analyzed.

To decide whether two words were semantically similar, we used the BERT language model (Devlin et al., 2018), that received a sentence as input and produced a list of word vectors as output, corresponding to each of the sentence's tokens. We chose BERT because it was an easily available state-of-the-art large language model at the time we performed this task. Using the vectors produced by BERT, we counted, for each snippet, how many words preceding the NC were semantically similar to the compound. That is, we calculated the cosine between each word in the snippet and each of the compound's component words and counted those whose cosine was larger than a certain threshold θ (see the next paragraph). Function words and words containing non-English characters were not considered (BERT word vectors for function words tend to be similar to a very large number of other vectors, since they appear in almost any kind of context). Figure 3 shows the results, i.e., the distribution of the number of similar words in a compound's snippet.

[5] This number was arbitrarily chosen; but reflected the fact that we intended to use these snippets in Experiment 1.

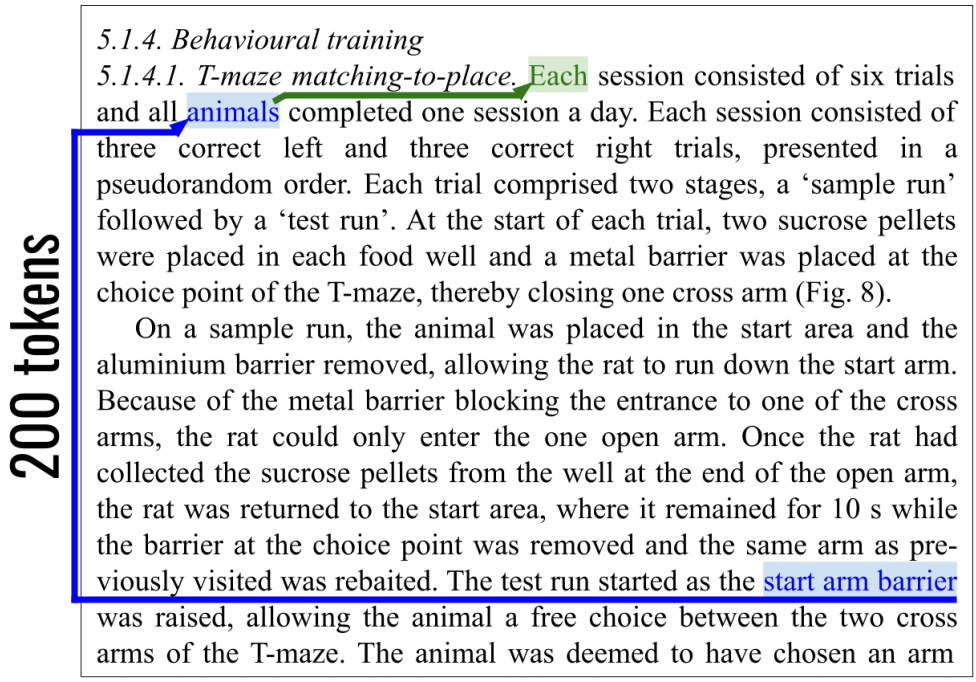


Figure 1. The procedure used to produce an NC snippet. From the NC, we counted 200 tokens back, landing in the word "animal" (blue). Then, we located the first token of the sentence that contained "animal" (green). (Adapted from Powell et al., 2017)

In order to find the best value for θ,[6] three annotators (all native or near-native English speakers) manually annotated the similar words in 20 NC snippets. These annotations were then used as a gold standard for the BERT annotation. A good θ should tag as "similar" roughly the same words as the gold standard. In addition, we noted that, on average 23.6% of the words in an NC context were annotated as similar to the NC. Thus, our choice of θ tried to mimic this average: we considered θ candidate values that would lead, on average, a certain number of snippet words (15%, 20%, 23.6%, 25%, 30% and 35%) to be considered similar to the NC. Among the candidates, $\theta = 0.503$ best mimicked the human annotation, and, on average, tagged 25% of the NC snippets as similar. Thus, a word in an NC context was treated as similar to the NC if the cosine of the angle between its vector and any of the NC vectors was bigger than 0.503.

[6] The θ selection procedure was complicated and is fully described in the supplementary materials.

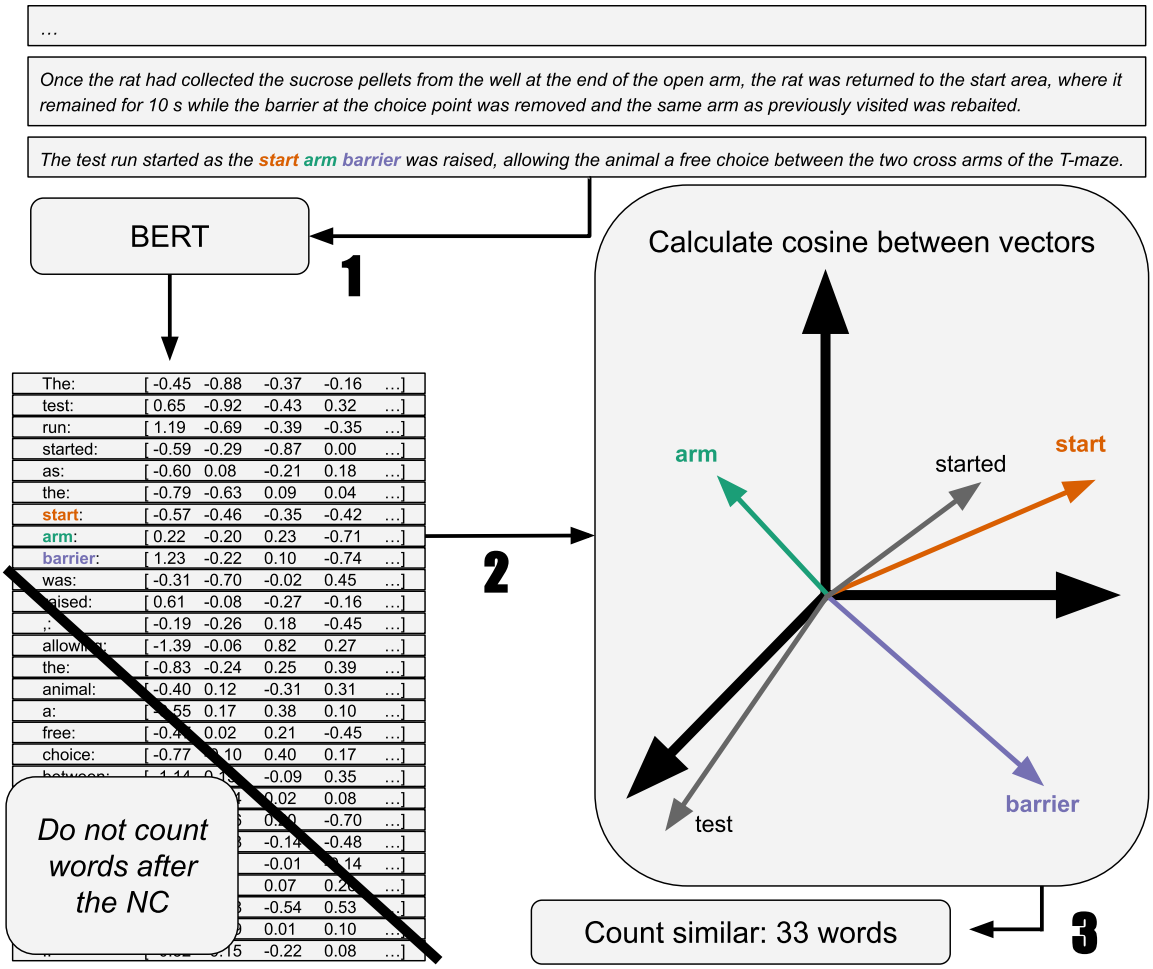


Figure 2. A sketch of the procedure used to calculate the similarity between the words in the NC snippet and the NC words. In step 1, each sentence of the snippet is fed into BERT, generating a vector (a list of numbers) for each word. These vectors can be interpreted as arrows pointing in a direction. In step 2, the cosine of the angle between these arrows and the NC words is calculated (the arrows shown in the image are fictitious: the real arrows cannot be represented in 3D). Similar words will point in a similar direction, producing cosines closer to 1; dissimilar words, pointing in different directions, will produce lower cosine values, up to -1. Finally, in step 3, we count the similar vectors. Note that, even though we feed BERT each sentence separately, we only count as similar words that *precede* the NC.

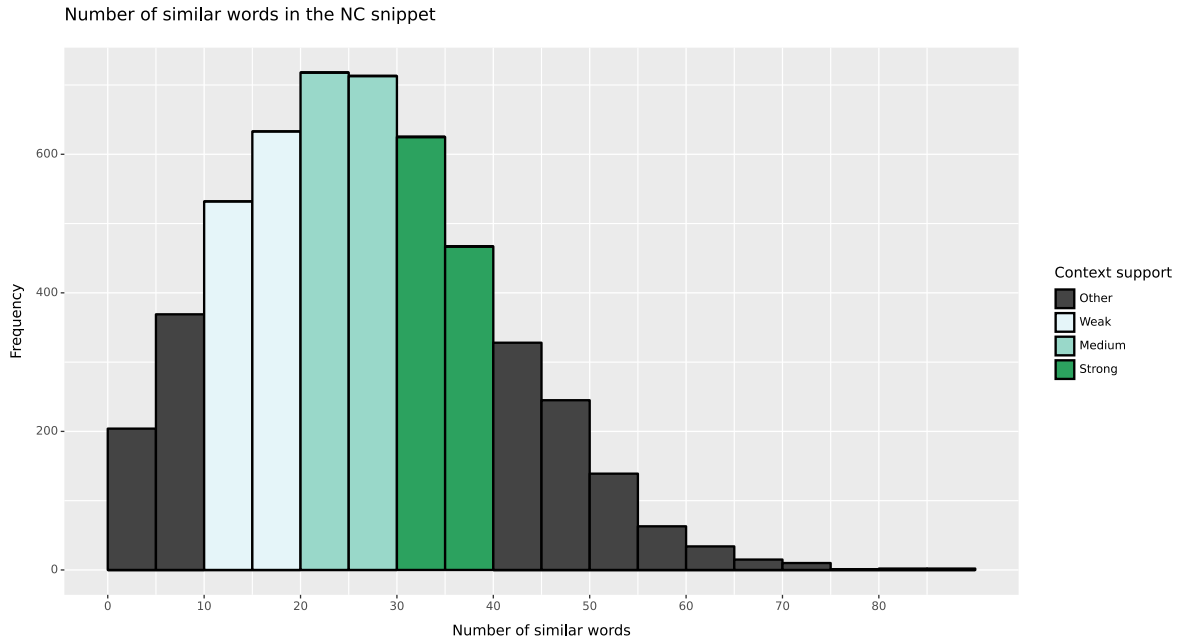


Figure 3. The distribution of presetting words preceding the corpus' NCs. The graph is Poisson distributed, as is typical of count data. The colors illustrate the categories used in Experiment 1.

As can be seen, there is a large variability in the number of presetting words immediately preceding an NC. We are now left with the question of whether this measure of contextual support can be empirically validated by behavioral measures of language processing. Thus, we ask whether NCs that receive more contextual support do lead to less processing difficulty.

Experiment 1

Participants

A total of 79 university students from the University of **<REDACTED>**, aged between 18 and 35 (Mean: 20.72, SD: 3.14), took part in the experiment for course credit. They were all English native speakers.

Materials

Given the distribution of Figure 3, we defined three categories of interest for the NCs. These ranges were arbitrarily defined, but chosen to avoid the extremes of the distribution.

- **Weak support**: the NC is preset by between 10 and 19 words
- **Medium support**: the NC is preset by between 20 and 29 words
- **Strong support**: the NC is preset by between 30 and 39 words

Table 2. *Total number of NCs in the categories of Experiment 1.*

	Weak support	**Medium support**	**Strong support**
3-word NCs	1093	1337	990
4-word NCs	72	82	91
Longer NCs	0	12	11

Table 2 shows the number of compounds in each category. From this pool, we selected 60 3-word compounds (20 in each category), and another 60 4-word compounds (20 in each category), roughly balanced by field of study. These 120 items were distributed into 6 lists of 20 items. Since a list of 20 items cannot have the same number of items for each condition, lists were constructed so that they contained at least 3 items of each condition (totalling 18 items), and then 2 additional items of two different conditions (see Figure 4a). We then used these 6 lists to produce 6 additional lists by merging them in pairs (see Figure 4b).

Procedure

We used the lists to create questionnaires where NCs were presented in context and in isolation (see Figure 5) and participants responded about the difficulty they perceived when reading them. There were a total of 12 questionnaires corresponding to the 12 lists. These questionnaires were fixed: All participants assigned to the same questionnaire saw the same questions in the same order. However, before creating the questionnaires, the lists were randomized, which ensured that most items appeared in two different positions in the two lists they were in.

Questionnaires were divided into three sections. In the first section, compounds were presented along with their context (4 practice trials, 20 critical trials; 3 catch trials[7] – see below) and participants answered two questions (Q1 and Q2 in Table 3). In the second section, compounds were presented in isolation (20 critical trials) and participants responded to one question (Q3). At the end of section 2, we also asked (QR) how much participants actually read the text snippets. Finally, in the third section,

[7] Out of the 23 critical trials, the trials 6, 11 and 15 were "catch trials" that were created in an attempt to determine whether participants were in fact paying attention to the questionnaire. Each catch trial was constructed by taking an NC from one field (e.g., Biology) and using it in a text snippet from another field (e.g., Linguistics). We initially expected the responses for Q2 in catch trials to be always a low value (say, smaller than 4), because participants would not be able to use the text passage in order to understand the NC, and therefore expected them to differ markedly from the Q2 in critical trials. Hence, we planned to exclude participants whose Q2 response in any catch trial was 5 or higher (criterion 1), or whose Q2 answer in critical trial was lower than 5 (criterion 2). After collecting the data, however, we noted that this approach would exclude 78 participants (out of 79). We therefore decided not to use this criterion in our analysis. Note, however, that this does not mean the participants were not reading any paragraphs. Looking better at the catch trials texts (see Appendix B), it is clear that the NCs were not as implausible as we initially expected: the paragraphs were composed of many technical terms that could be skipped by readers unfamiliar with the content. That is, even though it is true that some participants were not reading the whole paragraphs, this is not the reason why so many participants did not respond to the catch trials the way we expected. Indeed, many of them left thoughtful comments about the experiment in a final free text box at the end of the questionnaire suggesting they actually took it seriously. A comparison of Q2 in critical and catch trials is available in the Supplementary Materials.

participants answered simple questions about their language background (e.g., age, gender, education level, current study program, self-rated proficiency in the languages they speak, etc.).

Design

The experiment followed a within subjects 3x2x2 design (Context support x Compound length x Context/Isolation). Each participant saw all conditions presented in context, and all conditions presented in isolation, but all items presented in context were different from the items presented in isolation.

Analysis

Prior to analysis, participants who responded less than 5 to QR were excluded.

With the data of the remaining participants, we used the R programming language (R Core Team, 2024), and the *ordinal* R package (Christensen, 2023) to fit two Cumulative Linear Mixed Models (CLMM) using the *probit* link function (following Liddell & Kruschke, 2018). One of the models used the responses for Q1, and the other one used the responses for Q3. The probit link function is a Gaussian function with mean 0 and standard deviation 1. The *threshold coefficients* produced by the model are positions in this Gaussian that indicate the probability assigned by the model to a given response. For example, if the thresholds 1|2 and 2|3 are -0.5 and -0.1, respectively, then the probability (assigned by the CLMM) to the thresholds 1 and 2 are the areas under the curve from -∞ to -0.5, and from -0.5 to -0.1, respectively.

Both models were maximally specified (Barr et al., 2013).[8] Context support was sum-coded and ordered Weak, Strong, Medium, so that the results below show effects for the levels Weak and Strong; compound length was sum-coded.

Results

Out of the 79 participants, 16 (20.3%) responded less than 5 to QR and were therefore excluded from analysis.[9] The rest of the data was analyzed.

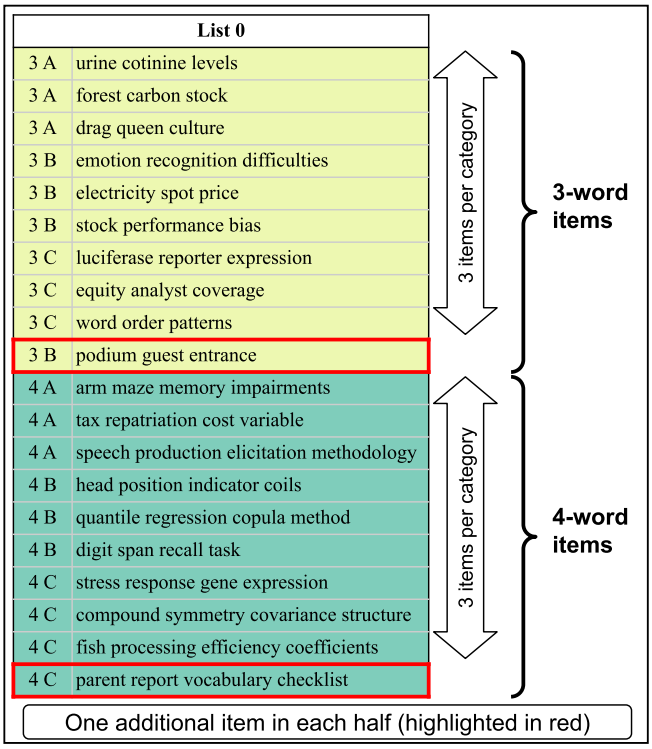
The in-context model analyzing Q1 yielded a main effect of compound length (b=-0.230, p=0.003), so that 4-word compounds led to higher response values than 3-word compounds (see Table A1). Figure 6a shows the distribution of the in-context responses for the different conditions.

Similarly, the in-isolation model analyzing Q3 yielded a main effect of length (b=-0.325, p<0.001) so that 4-word compounds led to higher response values than 3-word compounds (see Table A2). Figure 6b shows the distribution of the responses for the different conditions.

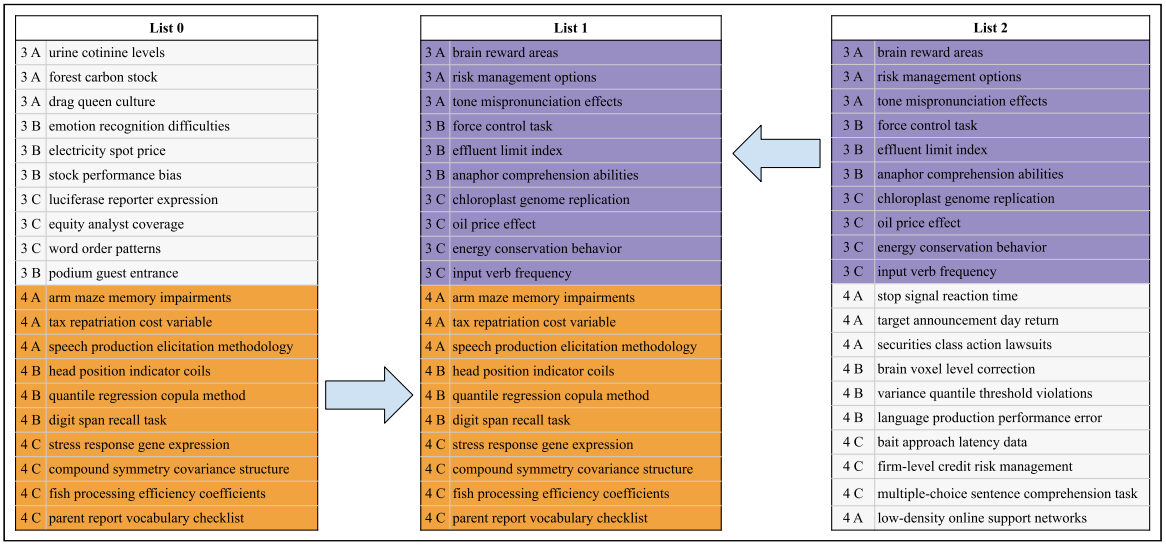
[8] The calls to clmm used the following formula:

```
answer ~ 1 + support*length + (1|item) + (1+support+length|participant)
```

[9] We also reran the models with all 79 participants, producing very similar results.

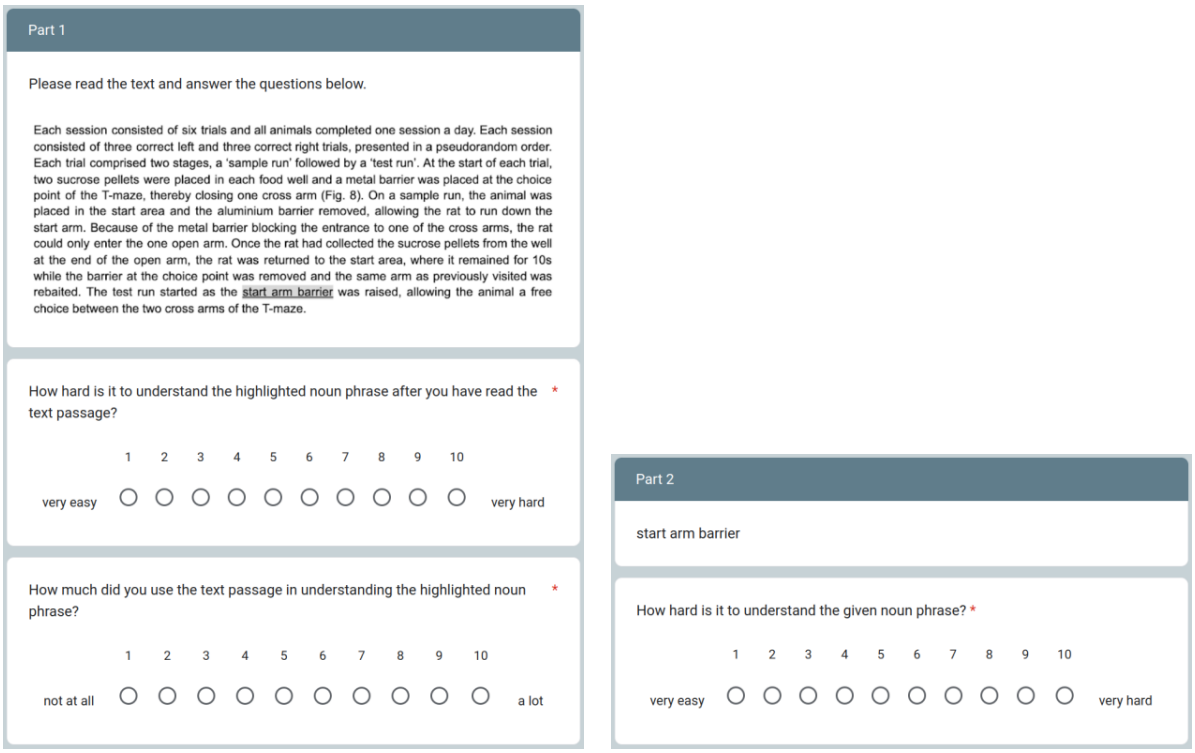


(a)



(b)

Figure 4. (a) One of the 6 non-overlapping item lists (these were numbered 0, 2, 4, 6, 8, 10) before it was randomized. (b) In order to produce the 6 additional lists (numbered 1, 3, 5, 7, 9, 11) we merged the first half of the next list with the second half of the previous one.



(a) (b)

Figure 5. Example questions for Section 1 (a) and 2 (b) of a questionnaire of Experiment 1. Items in Part 1 were different from those in Part 2 (so, a given participant who saw *start arm barrier* in Part 1 would not have seen it in Part 2, and vice-versa).

Table 3. *Questions responded to by the participants, which were considered in the analysis.*

		Experiment 1		Experiment 2	
Section	**Question**	**Question text**	**Answer**	**Question text**	**Answer**
1	Q1	How hard is it to understand the highlighted noun phrase after you have read the text passage?	Likert scale 1: very easy 10: very hard	After reading the text passage, how hard was it for you to understand the highlighted noun phrase?	Likert scale 1: very easy 10: very hard
	Q2	How much did you use the text passage in understanding the highlighted noun phrase?	Likert scale 1: very easy 10: very hard	Please provide a paraphrase for the highlighted noun phrase	Text response
2	Q3	How hard is it to understand the given noun phrase?	Likert scale 1: very easy 10: very hard	How hard is it to understand the given noun phrase?	Likert scale 1: very easy 10: very hard
	Q4	-	-	Please provide a paraphrase for the highlighted noun phrase.	Text response
End of 2	QR	In the first part of the experiment, how often did you actually read the *entire* text passages?	Likert scale 1: never 10: always	In the first part of the experiment, how often did you actually read the *entire* text passages?	Likert scale 1: never 10: always

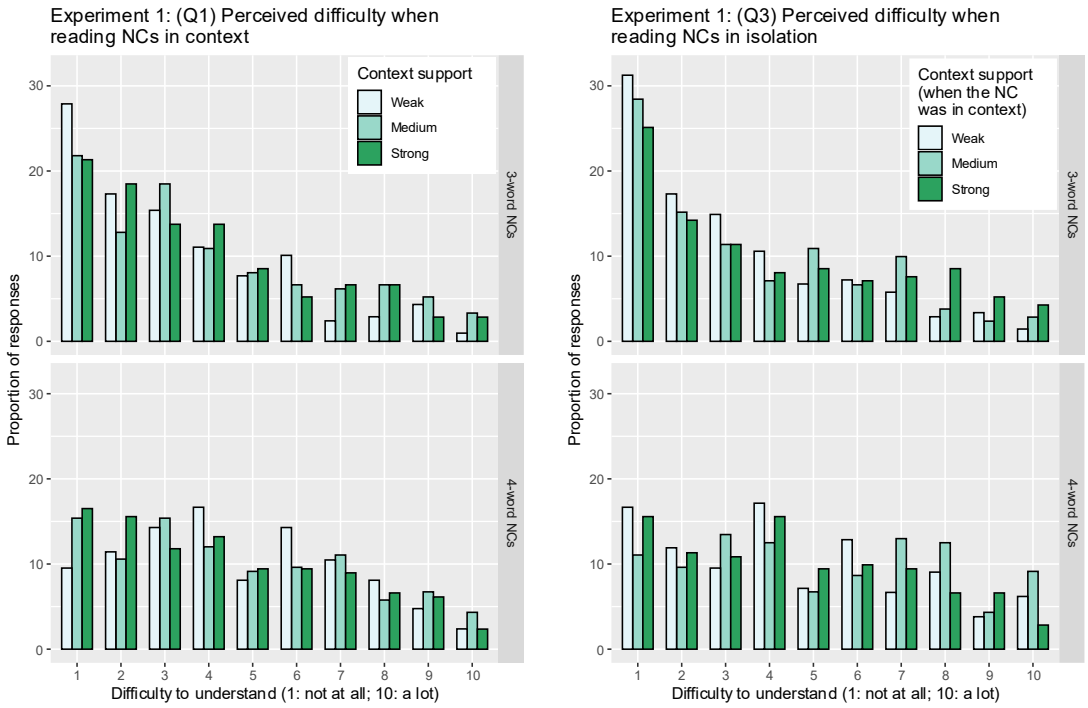


Figure 6. The distribution of Experiment 1 responses for Q1 (a) and Q3 (b).

Discussion

In this experiment, we aimed at testing the influence of contextual support on the processing of long NCs. In particular, we analyzed the self-rated difficulty perceived by readers when interpreting NCs presented either in context or in isolation. We manipulated length (3 or 4 words), and quality of contextual support (Weak, Medium, Strong) operationalized as the number of words preceding the NC that were similar to the NC words. This latter manipulation aimed at empirically validating our implementation of the counting measure proposed by Gamboa et al. (2024).

We expected longer NCs to cause more difficulty to readers. This is what we found: Length did lead to higher difficulty both when the NCs were presented in isolation and in context. In addition, we expected difficulty ratings to be significantly influenced by contextual support, at least for NCs presented in context, but no such influence was found either for NCs in context or in isolation. This may have to do with the fact that some of our items were quite familiar, composed of words that are often linked together (e.g., *heart rate variability*, *drag queen culture*, *control group participants*). Participants reading them would experience little difficulty regardless of the context, introducing noise into the data. Indeed, Table 4 shows the NCs rated as easiest and hardest when presented in context and in isolation for each condition. In Experiment 2, we extend Experiment 1 by manipulating NC familiarity.

Table 4. *The NCs rated as easiest in Q1 in Experiment 1.*

In Context					
Difficulty	**Support**	**3-word NC**	**Median**	**4-word NC**	**Median**
Easiest NCs	**Weak**	employment history information	1	short-term memory word span	1
		carbon reduction policies	1	stop signal reaction time	3
		drag queen culture	1	securities class action lawsuits	3
	Medium	health information content	1	heart rate variability changes	2
		podium guest entrance	1.5	world market oil price	2
		stock market anomalies	2	language production performance error	2
	Strong	start arm barrier	1	short-term object recognition memory	1
		word recognition errors	1	vehicle fuel economy ratings	1
		sentence repetition errors	1	multiple-choice sentence comprehension task	1
Hardest NCs	**Weak**	urine cotinine levels	4	reference memory probe trial	6
		forest carbon stock	4	digit span recall task	6
		tail suspension test	5	target announcement day return	6.5
	Medium	electricity spot price	4	eye movement artifact correction	6
		anaphor comprehension abilities	5.5	heat shock protein family	6
		head noun phrase	6.5	index fund ownership effect	7
	Strong	chloroplast genome replication	4.5	consensus analyst sales forecast	6
		commodity futures returns	5	compound symmetry covariance structure	7
		wh-question comprehension task	6	signal space separation method	7
In Isolation					
Difficulty	**Support**	**3-word NC**	**Median**	**4-word NC**	**Median**
Easiest NCs	**Weak**	stress response genes	1	heart rate perception test	2
		heart rate variability	1	heart rate variability index	2
		heart rate acceleration	1	press release disclosure measures	2
	Medium	control group participants	1	world market oil price	1
		stock market anomalies	1	gender identification accuracy data	2
		eye movement data	1	heart rate variability changes	2
	Strong	response times analyses	1	vehicle fuel economy ratings	1
		word recognition errors	1	multiple-choice sentence comprehension task	1
		sentence repetition errors	1	infant speech perception findings	2
Hardest NCs	**Weak**	forest carbon stock	4.5	arm maze memory impairments	6.5
		bottleneck approval structure	4.5	speech production elicitation methodology	6.5
		cell surface expression	5	digit span recall task	7
	Medium	force control task	5	repatriation tax cost variable	7
		core clock molecule	6	index fund ownership effect	8
		response reversal stages	7	variance quantile threshold violations	8
	Strong	equity analyst coverage	6.5	bait approach latency data	7
		referent selection trials	6.5	ingenuity pathway knowledge base	7
		start arm barrier	7	signal space separation method	7

Experiment 1 had two other limitations. First, we note that it is not really clear, based solely on self-rated difficulties, whether the participants *truly* understood the NCs, especially if the NC was not among the familiar ones discussed in the previous paragraph. An NC rated as "easy" may very well have been incorrectly understood. To address this issue, in Experiment 2 we removed Q2 and instead asked participants to paraphrase the NC in both sections of the questionnaire.

The second limitation has to do with the NC contexts we selected. As Gamboa et al. (2024) suggested, when an NC appears for the first time in a paper, it is typically well introduced by its context.

But once used, it can be reused without necessarily being thoroughly reintroduced. In Experiment 1, we did not ensure that the NC contexts (which were extracted from real scientific articles) contained only the first occurrence of a given NC in a paper. This is ensured in Experiment 2.

Experiment 2

In Experiment 2 we repeated Experiment 1 addressing some of its limitations. Participants filled out questionnaires where they rated the difficulty of and produced paraphrases for familiar and unfamiliar NCs presented either in context or in isolation.[10] Contexts were shortened to about 100 words[11] in order to encourage reading, and only first-usage NCs were selected. We simplified the experiment design by only considering 3-word NCs and only selecting NCs from the categories "Weak" and "Strong".

In order to decide whether an NC was familiar or not, we calculated a familiarity score based on the frequency of the NC (using the Google Ngram viewer; Michel et al., 2010) and the existence of Wikipedia pages referring to the NC or to its subparts (more details can be found in the supplementary materials).

Participants

A total of 79 university students from the University of **<REDACTED>**, aged between 17 and 29 (Mean: 20.5, SD: 2.64), took part in the experiment for course credit. They were all English native speakers.

Materials

Similarly to the items of Experiment 1, contexts were constructed with the procedure of Figure 1, but counting 100 tokens back from the NC (instead of 200). We then arbitrarily selected 96 NCs along with their contexts from the corpus and arranged into 4 lists of 24 items. Each list contained an equal number of familiar and unfamiliar NCs, with Weak and Strong contextual support, forming a 2x2 design. In addition, each list had an equal number of items from Biology, Economics and Linguistics. Four additional lists were produced using the procedure depicted in Figure 4b.

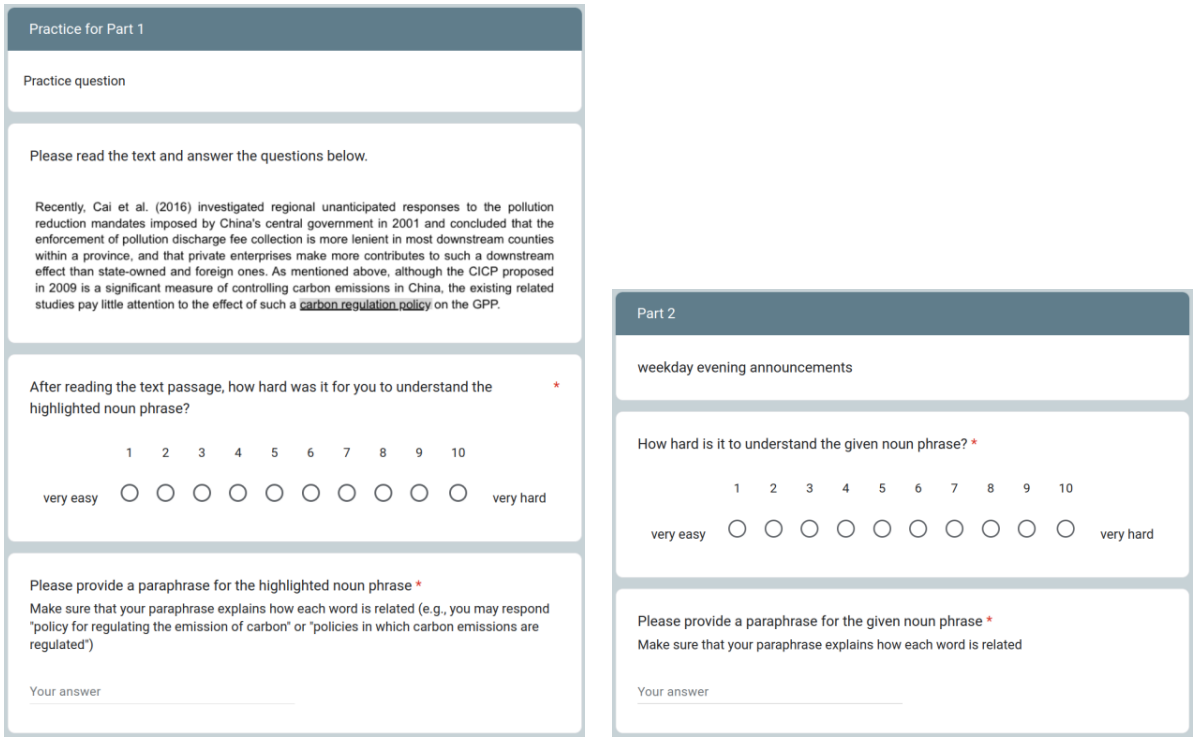
Procedure

The procedure was similar to that of Experiment 1. Participants answered questionnaires containing NCs presented in context and in isolation, divided into the same 3 sections. There were a total of 8 questionnaires corresponding to the 8 lists, which were randomized before questionnaire construction and presented always in the same order. The phrasing of the questions was slightly modified, and Q2 was replaced by a paraphrasing task (see Table 3).

In order to encourage participants to produce clear paraphrases, we provided examples of paraphrases for the practice items (see Figure 7a). In addition, Q2 and Q4 displayed the cue (in smaller font) "Make sure that your paraphrase explains how each word is related".

[10] A first version of Experiment 2 had participants see the same list of items both in context and in isolation, and in a non-randomized order. Since these characteristics may have affected the results of the experiment, we do not report on those results.

[11] With shorter contexts, we recounted the number of similar words in the (new) contexts for all NCs and recategorized NCs according to the three contextual support categories used in Experiment 1. This recategorization was made so that the amount of NCs in each category remained the same (see Supplementary Materials).



(a) (b)

Figure 7. Example questions for Section 1 (a) and 2 (b) of a questionnaire of Experiment 2. The example in Section 1 is a practice trial, showing paraphrase suggestions.

Design

The experiment followed a within subjects 2x2x2 design (Context support x Familiarity x Context/Isolation). Each participant saw all conditions presented in context, and all conditions presented in isolation, but all items presented in context were different from the items presented in isolation.

Analysis

Prior to analysis, participants who responded less than 5 to QR were excluded.

Three annotators (two L1 English speakers) independently tagged the paraphrases produced by all participants (Q2 and Q4). These tags were used as votes and each paraphrase was assigned the tag that received the most votes. Annotators could tag the paraphrases as *correct*, *incorrect but plausible*, *incorrect*, or *weird* (the latter indicating that the participant did not engage in the task – e.g., paraphrases such as "I don't know" or "X"). In addition, annotators could use the tag *unsure* to renounce their vote. Then, the three discussed their votes for all tied cases (e.g., each annotator used a different tag for the paraphrase; or all annotators used the tag *unsure*), and assigned a final tag to the paraphrase according to their discussion. We used these tags to calculate accuracy for each of the conditions.

In order to analyze Q1 and Q3 (the difficulty perceived when reading the NCs), items whose final paraphrase (Q2 and Q4, respectively) was not *correct* were discarded. The remaining items were then fed

(as in Experiment 1) into two maximally specified CLMM models:[12] one for the NCs in context and another for the NCs in isolation. Both Context support and Familiarity were sum coded.

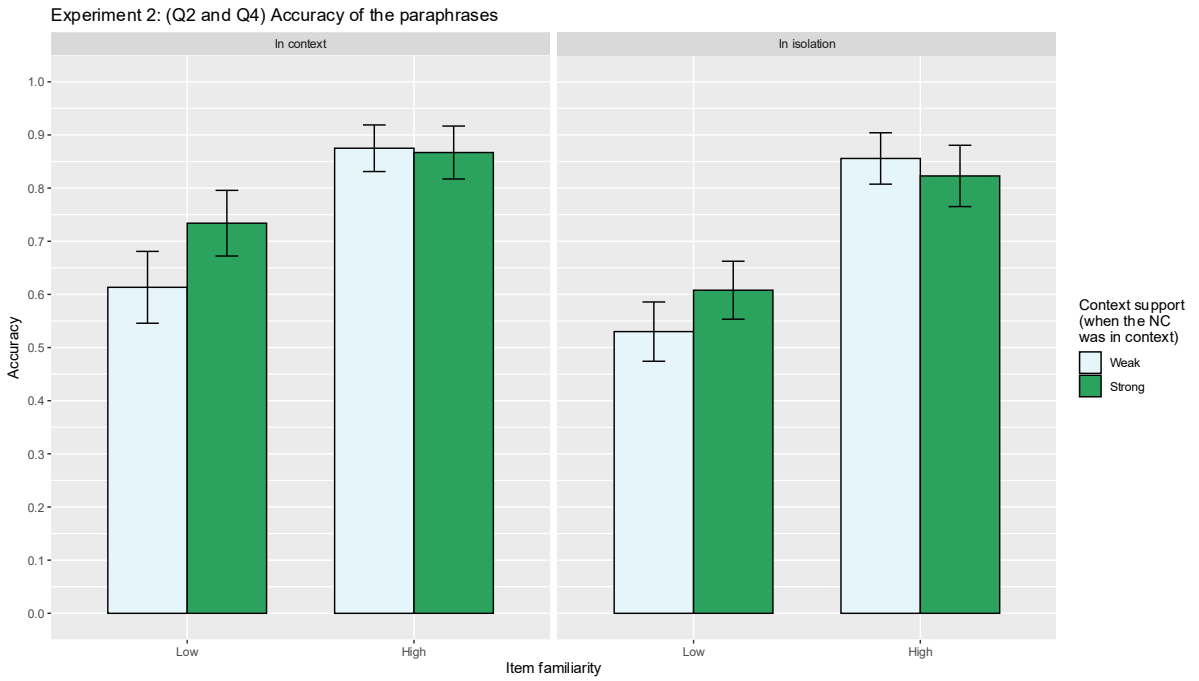


Figure 8. The proportion of paraphrases tagged as *correct* for each condition in Experiment 2. Error bars indicate confidence intervals.

Results

Out of the 79 participants, 27 (34.1%) responded less than 5 to QR and were therefore excluded from analysis. The rest of the data (2496 trials) was analyzed.

Figure 8 shows the proportion of paraphrases tagged as *correct* by the annotators when the NCs were presented in context (Q2) and in isolation (Q4). A total of 681 (27.3%, 297 in context, 384 isolated) responses were not tagged as *correct* and were therefore excluded from the subsequent analysis.[13]

The remaining answers (951 in context, 864 in isolation) were used to fit the two models. The in-context model (Table A3) analyzing Q1 yielded a main effect of familiarity ($b = 0.694$, $p < 0.001$), so that more familiar NCs were rated easier than less familiar ones. In addition, it yielded a significant interaction between context support and familiarity ($b = 0.167$, $p = 0.033$), indicating a larger difference in difficulty ratings for low familiarity items than for high familiarity items. Figure 9a shows the distribution of the in-context responses for the different conditions.

The in-isolation model (see Table A4) analyzing Q3 also yielded a main effect of familiarity ($b = 0.808$, $p < 0.001$) so that more familiar NCs were rated easier than less familiar ones. Figure 9b shows the distribution of the responses for the different conditions.

[12] The calls to clmm used the following formula:

```
answer ~ 1 + support*familiarity + (1|item) +
(1+support+familiarity|participant)
```

[13] Given the large data losses, we additionally ran models (1) without excluding participants, and (2) without excluding either participants or incorrect trials. In both cases, results were very similar to those reported here.

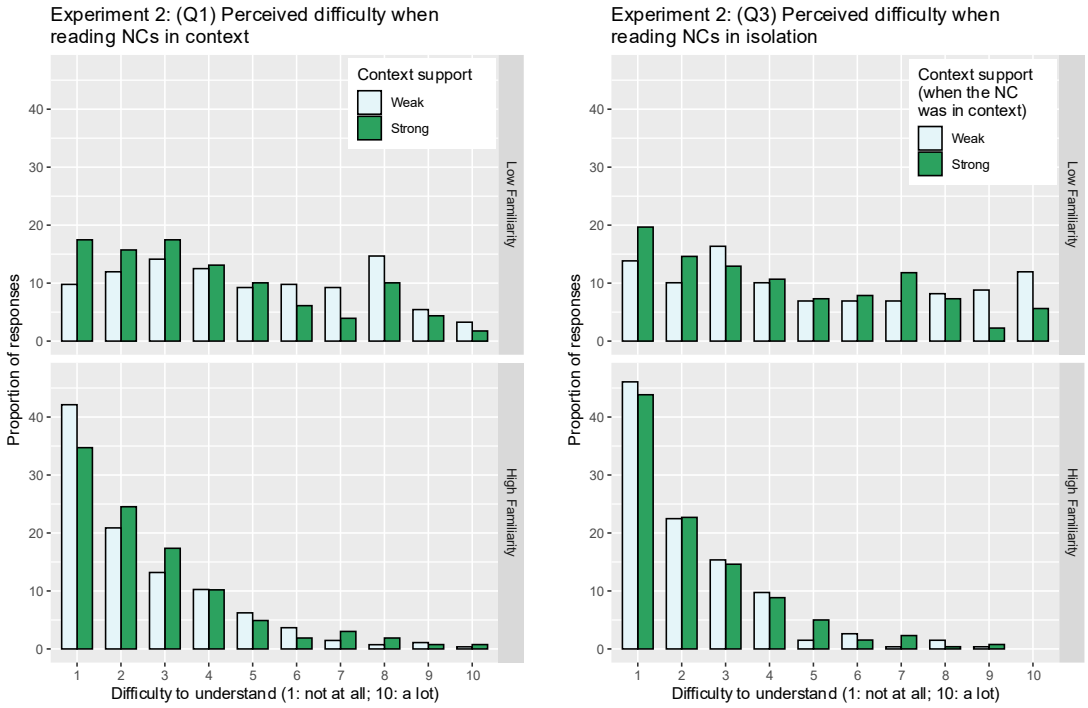


Figure 9. The distribution of Experiment 2 responses for Q1 (a) and Q3 (b).

Discussion

In Experiment 2, we ran a similar procedure to Experiment 1, addressing some of its limitations, and trying to find evidence suggesting that contextual support, as operationalized through the similar-words measure, influences the difficulty perceived by participants when reading NCs.

As expected, both models showed a clear effect of familiarity: Participants produced more accurate paraphrases for high familiarity NCs, and judged them easier than low familiarity ones. When it comes to contextual support, as expected, the significant interaction between familiarity and support found for the NCs presented in context did indicate that low familiarity NCs are much more influenced by the context than high familiarity ones. Interestingly, this interaction was not found for NCs presented in isolation.

General Discussion

In this study, we had one main focus: to understand how (much) the preceding context affects the processing of nominal compounds (NCs). Exactly what we mean by context, however, is difficulty to quantify. In order to represent it, we turned to the similar-words measure proposed in Gamboa et al. (2024b). Thus, associated with this measure, we defined two goals. Our first goal was to implement it, highlighting the practical considerations involved in such an effort, and evaluating the use of such a measure in psycholinguistic experiments, i.e., verifying whether their results conform to our expectations, and exploring how or whether this measure can be useful in future studies. Assuming that it *is* useful, and

in an attempt to partially explain why long NCs are used in scientific texts, our second goal was to use this measure to investigate how the context affects the difficulty perceived by participants when reading long NCs with 3 or more words, thus exploring how or whether it advances our knowledge on how NCs are comprehended.

We implemented the similar-words measure and applied it on a pool of 5100 NCs. In order to implement it, we used BERT to produce word vectors for each NC, as well as for all words in the vicinity of the NC. Using these vectors, the similar-words measure output was a count of how many words (vectors) in the NC vicinity were similar to the NC itself (Figure 3).

We then selected a number of NCs from this pool to use as items in Experiments 1 and 2, dividing them into three categories (Weak, Medium and Strong contextual support). We expected shorter NCs, as well as NCs with more contextual support, to lead to lower difficulty judgments. The results initially (Experiment 1) did not follow our expectations. While there was a clear effect of length (4-word NCs were generally judged harder then 3-word NCs), we found no effect of context support, neither for NCs presented in context nor in isolation.

We found clearer results after controlling for item familiarity (Experiment 2). The significant interaction for NCs presented in context indicated that, while high familiarity NCs were not judged very different in the different categories, low familiarity NCs did show a substantial difference between weak and strong contextual support. However, this interaction was not present in the model we fit for the NCs in isolation. It seems, therefore, that it is not the case that well-supported NCs *are well-supported by their context* because *they require support*, as alluded to in the introduction. Rather, the fact that even NCs that are not particularly difficult receive a substantial amount of contextual support may help explain why NCs are so frequent in scientific papers. An NC is not typically a complicated concept coined ad hoc, but a natural shorthand term used to succinctly express concepts that are already clear from the context.

How NC difficulty compares for NCs presented in context vs. in isolation

In Experiments 1 and 2, we did not compare the difficulty perceived by participants when reading NCs in context vs. in isolation. Participants always saw the NCs presented in context first, in Part 1 of the questionnaire, and only then proceeded to Part 2, where they read NCs in isolation. Here, we attempt a post-hoc comparison of the two presentation modes: For each experiment, we use the two models we fit for NCs presented in context (i.e., for the Q1 responses) and in isolation (i.e., for the Q3 responses).

We follow the ideas from Liddell and Kruschke (2018), in their analyses of ordinal data. Given the thresholds produced by their ordinal model, they derived a "latent mean" for the model, i.e., a continuous value (in principle, ranging from $-\infty$ to $+\infty$) indicating the mean of the latent distribution from which the responses would be sampled.[14] This was then used to produce estimated responses by item in this latent domain (e.g., an item with latent mean 3 would be easier than an item with latent mean 8). We adapt their procedure to produce latent mean values for each of the items in both experiments.

Figure 10a shows the calculated latent mean for all items in Experiment 1. As can be seen, besides the clear effect of length discussed earlier, there is a clear correlation between the difficulty experienced for items in context and in isolation: items that were rated as harder to comprehend in context were typically also rated as harder in isolation. Similar results can be seen in Figure 10b. Besides the clear effect of familiarity, it is possible to see how, on average, weakly supported items were rated as a bit

[14] Their discussion is based on a Bayesian model, but this specific conversion is agnostic to the type of model being used.

harder than strongly supported items. (Interestingly, the graph suggests that there could also have been an interaction for the model in isolation, but this interaction did not reach significance – see Table A4.)

We stress that this analysis should be taken with caution. Given the high number of items in both experiments, each item received a very small number of responses (on average, around 10.5 and 9.5 responses per item for each model of Experiment 1 and Experiment 2, respectively). Hence, the variability per item was large, and could not be compensated by the random slopes by item used in the CLMMs, causing the model to not fit the data very well. While we do not believe that this would have affected the aggregate results (the main results of Experiments 1 and 2 reported above), we would recommend caution in interpreting per item values.

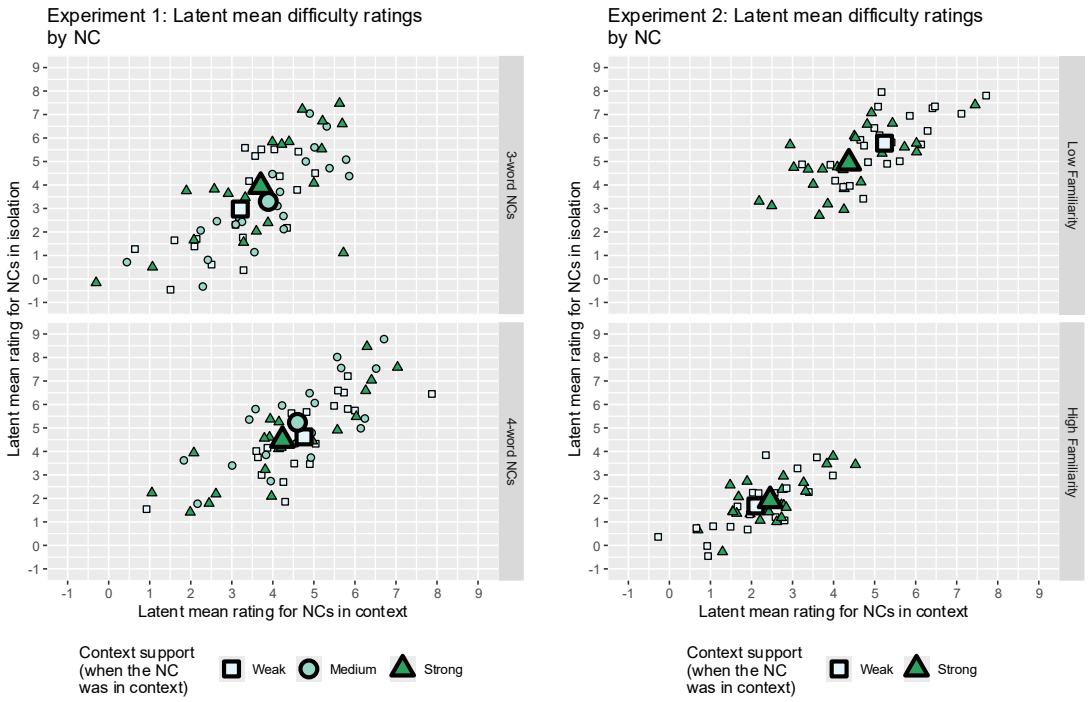


Figure 10. The latent mean of all items in Experiments 1 (a) and 2 (b), in context and in isolation. The highlighted points indicate the average latent mean of all items of a given condition.

Limitations, possible improvements, and future uses of the similar-words measure

It should be noted that the similar-words measure, the way it is implemented here, has some limitations. First, it does not take into account the structure of the sentences preceding the NC. All it does is to count words, and the decision on whether the words are used in a way that is related to the NC meaning is left to the language model (here, BERT). For example, in (3), an ideal language model should probably treat *fire* and *firing* as different words, even if they share the same root, because the sense and the structure of the sentences suggest completely different meanings. We argue this is not a severe limitation, given the advances in language modeling in the last few years (e.g., BERT did produce dissimilar vectors for the two words).

(3) *The catastrophic damage of these extreme weather events in the region has led to upper management layoffs, especially in the case of companies located in communities affected by forest* ***fire****. Company board minutes often mention such external events as relevant to their CEO* ***firing*** *decision.*

Second, some words may not be present in BERT's vocabulary. This was the case for many words in our items, such as *idiosyncratic*, *siting*, *untabulated*, *CEO*, *IPO*, *SPAC*, etc. In these cases, BERT will still produce a word vector, but it will not necessarily correspond well to the word. Similarly, some words may be "hidden" in acronyms and abbreviations, which BERT will likely ignore even if a real reader would have no trouble in realizing the acronym meanings. This was the case for words such as *ROI* (containing *return* and *investment*), *OPEC* (containing, e.g., *petroleum*) or *IPO* (i.e., *initial public offering*).

Finally, even though it is not the case with BERT, the tokenization of many modern language models does not map *words* into *tokens* in the way depicted in Figure 2 (e.g., a word such as *bicycle* may be sometimes tokenized as *b* and *icycle*; but note that this often ends up usefully breaking words such as *globalization* into *global* and *ization*). While it is possible to blindly use similar-tokens measure with this kind of tokenization, the measure partially loses its intuitive appeal.

We believe, however, that these limitations are not serious enough to render the similar-words measure useless. Indeed, implementations of the similar-words measure using more recent language models may improve performance, and allow for more context to be taken into account. That is, instead of inputting each sentence separately to BERT, newer language models may be able to handle the paragraph as a whole, generating higher-quality word vectors.

In addition, alternative versions of the measure may be able to completely avoid the empirical threshold selection process we followed in this paper. We only needed a threshold because we needed to make a binary decision on whether two words were similar or not. If, however, the similarity values between the words were always positive (as was surprisingly often the case with BERT), then, instead of counting the number of similar words, we could have just summed the similarity values of the words preceding the NC. This would allow for a continuous measure and better distinguish between words that are completely different from the NC components from words that fell just a bit short of the threshold.

In summary, Experiments 1 and 2 show how the similar-words measure implemented here does capture some important properties of the support provided to a given NC by its context. We hope that this measure may be used in future studies trying to quantify contextual support.

Conclusion

In two experiments, we investigated the difficulty perceived by readers of long NCs, and how the context influences this difficulty. Our results showed that the contextual support received by NCs does not depend on how difficult they are to read in isolation. In addition, after controlling for familiarity, we found that the comprehension difficulty associated with unfamiliar NCs is more affected by the context than that associated with familiar ones.

These experiments were based on items that manipulated the support given to NCs by their preceding context. In order to quantify this contextual support, we implemented a measure based on how

many words in the NC context are similar to the NC. We hope that the experiments reported here may be useful inspiration for future studies seeking to measure contextual support.

Acknowledgements

Since the beginning of this project, we have received help from a number of people who have contributed on many fronts, from programming some of the scripts, to extracting the NCs from the corpus, to reviewing the experimental items, all the way to painstakingly annotating the paraphrases produced by the participants in Experiment 2. None of these tasks would have been possible without the hard work of Abigail Hodge, Hannah Lee, Luc Henriquez, Marion Anglin, Mark Murphy, and Mateo Vargas-Nuñes. We would like to also thank Christopher Allison, Fenia Karkaletsou, Laís Muntini, Leigh Fernandez, Omar Jubran, Rhiannon Stewart, Tatiana Pashkova, and Xin Ying Lee for their comments on a previous draft of this paper. Finally, we are grateful for the feedback provided by the audience of the 15th Biannual Conference of the German Cognitive Science Society (KogWis 2022), in Freiburg, Germany, and of the 12th Mental Lexicon Conference 2022, in Niagara-on-the-Lake, Canada, where parts of this study was presented.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255-278. https://doi.org/10.1016/j.jml.2012.11.001

Benjamin, S., & Schmidtke, D. (2023) Conceptual combination during novel and existing compound word reading in context: A self-paced reading study. *Memory & Cognition, 51*, 1170–1197. https://doi.org/10.3758/s13421-022-01378-z

Bauer, L., Lieber, R., & Plag I. (2013) *The Oxford Reference Guide to English Morphology*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198747062.001.0001

Biber, D., & Gray, B. (2011). Grammatical change in the noun phrase: The influence of written language use. *English Language & Linguistics*, *15* (2), 223–250. https://doi.org/10.1017/S1360674311000025

Bürkner, P. C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, Vol. 2(1), 77–101. https://doi.org/10.1177/2515245918823199

Cohen, A. L., & Staub, A. (2014). Online processing of novel noun–noun compounds: Eye movement evidence. *Quarterly Journal of Experimental Psychology*, *67*(1), 147–165. https://doi.org/10.1080/17470218.2013.796398

Christensen R. (2023). *ordinal: Regression Models for Ordinal Data* (Version 2023.12-4) [R package]. Retrieved from https://CRAN.R-project.org/package=ordinal

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gagné, C. L., & Spalding, T. L. (2009). Constituent integration during the processing of compound words: Does it involve the use of relational structures? *Journal of Memory and Language*, *60*, 20–35. https://doi.org/10.1016/j.jml.2008.07.003

Gagné, C. L., & Spalding T. L. (2013). Chapter Three - Conceptual Composition: The Role of Relational Competition in the Comprehension of Modifier-Noun Phrases and Noun–Noun Compounds. In B. H. Ross (Ed.) *Psychology of Learning and Motivation* (pp. 97–130). Amsterdam: Academic Press. https://doi.org/10.1016/B978-0-12-407187-2.00003-4

Gamboa, J. C. B., Fernandez, L. B., & Allen, S. E. M. (2024a). Investigating the Uniform Information Density hypothesis with complex nominal compounds. *Applied Psycholinguistics, 45*(2), 322–367. https://doi.org/10.1017/S0142716424000092

Gamboa, J., Braun, K., Järvikivi, J. & Allen, S. (2024b). The distributional properties of long nominal compounds in scientific articles: An investigation based on the Uniform Information Density Hypothesis. *Corpus Linguistics and Linguistic Theory*. https://doi.org/10.1515/cllt-2023-0028

Gamboa, J., Fernandez, L. B., & Allen, S. E. M. (2025). Investigating the Uniform Information Density Hypothesis in L2 with complex nominal compounds. [Manuscript in preparation].

Geer, S. E., Gleitman, H., & Gleitman, L. (1972). Paraphrasing and remembering compound words. *Journal of Verbal Learning and Verbal Behavior*, *11* (3), 348–355. https://doi.org/10.1016/S0022-5371(72)80097-5

Gleitman, L. R., & Gleitman, H. (1970). *Phrase and paraphrase: Some innovative uses of language.* Norton.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strenth Natural Language Processing in Python (Version 2.1.6) [Computer Software]. Retrieved from https://spacy.io

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, *61*(1), 23–62. https://doi.org/10.1016/j.cogpsych.2010.02.002

Kvam, A. M. (1990). Three-part noun combinations in English, composition – meaning – stress. *English Studies: A Journal of English Language and Literature*, *71* (2), 152–161. https://doi.org/10.1080/00138389008598684

Kuperman, V., Bertram, R., & Baayen, R. H. (2008). Morphological dynamics in compound processing. *Language and Cognitive Processes*, *23* (7), 1089–1132. https://doi.org/10.1080/01690960802193688

Levy R. (2008) Expectation-based syntactic comprehension. *Cognition*, *106* (3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Limaye, M., & Pompian, R. (1991). Brevity versus clarity: The comprehensibility of nominal compounds in business and technical prose. The Journal of Business Communication, 28 (1), 7–21. https://doi.org/10.1177/002194369102800102

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong?. *Journal of Experimental Social Psychology*, *79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Michel, J.B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., The Google Books Team, Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A., Aiden, E.L. (2010). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science 331*(6014), 176–182. https://doi.org/10.1126/science.1199644

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C.J. Burges and L. Bottou and M. Welling and Z. Ghahramani and K.Q. Weinberger (Eds.), *Proceedings of the 26th International Conference on*

Neural Information Processing Systems. (pp. 3111–3119). Red Hook, NY: Curran Associates. https://proceedings.neurips.cc/paper/2013

Montero, B. (1996). Technical communication: Complex nominals used to express new concepts in scientific English-causes and ambiguity in meaning. The ESPecialist, 17 (1), 57–72. Retrieved from https://revistas.pucsp.br/index.php/esp/article/view/9476/7042

Nakov, P. (2013). On the interpretation of noun compounds: Syntax, semantics, and entailment. *Natural Language Engineering*, *19* (3), 291–330. https://doi.org/10.1017/S1351324913000065

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, *15*, 625–632. https://doi.org/10.1007/s10459-010-9222-y

Powell, A. L., Nelson, A. J., Hindley, E., Davies, M., Aggleton, J. P., & Vann, S. D. (2017). The rat retrosplenial cortex as a link for frontal functions: A lesion analysis. *Behavioural Brain Research*, *335*, 88–102. https://doi.org/10.1016/j.bbr.2017.08.010

R Core Team (2024). *R: A Language and Environment for Statistical Computing* (Version 4.4.0) [Computer Software]. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, *11*, 95–130. https://doi.org/10.1613/jair.514

Robitzsch, A. (2020). Why Ordinal Variables Can (Almost) Always Be Treated as Continuous Variables: Clarifying Assumptions of Robust Continuous and Ordinal Factor Analysis Estimation Methods. *Frontiers in Education*, Vol. 5. https://doi.org/10.3389/feduc.2020.589965

Sahlgren, M. (2008). The distributional hypothesis. *Rivista di Linguistica (Italian Journal of Linguistics)*, *20*(1), 33–53. https://linguistica.sns.it/RdL/20.1/Sahlgren.pdf

Schmidtke, D., Van Dyke, J. & Kuperman, V. (2021). CompLex: An eye-movement database of compound word reading in English. *Behavior Research Methods*, *53*, 59–77. https://doi.org/10.3758/s13428-020-01397-1

Sikos, L., Greenberg, C., Drenhaus, H., & Crocker, M. W. (2017). Information density of encodings: The role of syntactic variation in comprehension. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 3168–3173).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon and U. Von Luxburg and S. Bengio and H. Wallach and R. Fergus and S. Vishwanathan and R. Garnett (Eds.), *Proceedings of the 31st International Conference on Neural Information Processing Systems*. (pp. 5998–6008). Red Hook, NY: Curran Associates. https://papers.nips.cc/paper_files/paper/2017

Appendix A
Full Model Outputs

Experiment 1

Table A1. *Experiment 1 model results for the question "How hard is it to understand the highlighted noun phrase after you have read the text passage?" (Q1), with compounds presented in context.*

	Estimate	Std. Error	z value	Pr(>|z|)
Support 1	-0.040	0.110	-0.366	0.714
Support 2	-0.048	0.108	-0.442	0.658
Length 1	-0.230	0.077	-2.986	0.003**
Support 1: Length 1	-0.153	0.108	-1.417	0.156
Support 2 : Length 1	0.099	0.108	0.913	0.361
Threshold coefficients				
1|2	-1.274	0.114	-11.185	
2|3	-0.619	0.110	-5.634	
3|4	-0.055	0.109	-0.503	
4|5	0.417	0.109	3.807	
5|6	0.747	0.111	6.759	
6|7	1.156	0.113	10.242	
7|8	1.574	0.117	13.486	
8|9	2.031	0.123	16.456	
9|10	2.693	0.142	18.927	

Table A2. *Experiment 1 model results for the question "How hard is it to understand the given noun phrase?" (Q3), with compounds presented in isolation.*

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
Support 1	-0.140	0.137	-1.026	0.305
Support 2	0.055	0.135	0.408	0.683
Length 1	-0.325	0.096	-3.378	0.001***
Support 1: Length 1	-0.066	0.135	-0.488	0.625
Support 2 : Length 1	0.199	0.135	1.475	0.140
Threshold coefficients				
1|2	-1.244	0.122	-10.167	
2|3	-0.610	0.119	-5.127	
3|4	-0.115	0.118	-0.977	
4|5	0.353	0.119	2.983	
5|6	0.695	0.120	5.819	
6|7	1.093	0.121	9.006	
7|8	1.564	0.125	12.516	
8|9	2.091	0.132	15.897	
9|10	2.584	0.142	18.174	

Experiment 2

Table A3. *Experiment 2 model results for the question "After reading the text passage, how hard was it for you to understand the highlighted noun phrase?" (Q1), with compounds presented in context.*

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
Support	0.073	0.080	0.916	0.360
Familiarity	0.694	0.091	7.587	< 0.001***
Support: Familiarity	0.167	0.078	2.136	0.033*
	Threshold coefficients			
1|2	-1.132	0.150	-7.539	
2|3	-0.267	0.147	-1.825	
3|4	0.389	0.147	2.655	
4|5	0.909	0.149	6.109	
5|6	1.306	0.152	8.612	
6|7	1.625	0.155	10.471	
7|8	1.947	0.160	12.147	
8|9	2.665	0.178	14.937	
9|10	3.293	0.208	15.846	

Table A4. *Experiment 2 model results for the question "How hard is it to understand the given noun phrase?" (Q3), with compounds presented in isolation.*

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
Support	0.070	0.081	0.860	0.390
Familiarity	0.808	0.083	9.692	< 0.001***
Support: Familiarity	0.123	0.078	1.576	0.115
Threshold coefficients				
1|2	-0.951	0.142	-6.686	
2|3	-0.186	0.140	-1.328	
3|4	0.442	0.140	3.151	
4|5	0.922	0.143	6.463	
5|6	1.209	0.145	8.332	
6|7	1.503	0.149	10.117	
7|8	1.886	0.155	12.180	
8|9	2.293	0.165	13.924	
9|10	2.704	0.180	15.006	

Appendix B
Catch trials of Experiment 1

The following were the three catch trials used in Experiment 1.

These results, along with those in Table 2, offer preliminary evidence that aggregate earnings convey monetary policy news that can contribute to the negative aggregate earnings-returns association. Analyzing the role of monetary policy news at the quarterly frequency presents several concerns that can be mitigated by short-window tests. In this section, we discuss our short-window tests, in which we employ both monthly and FOMC announcement-day measures of returns and surprises. We begin with monthly analyses for the 1989–2007 period. Table 3 reports descriptive statistics for monthly aggregate earnings changes, the expected and surprise components of changes in the target rate, and macroeconomic forecast errors. We again partition our sample into periods of “Decrease”, “Increase” or “No Change”. The monthly descriptive statistics are consistent with our quarterly numbers. In particular, we find that the target rate decreases (increases) in 18.4% (13.9%) of the sample months and, as expected, average and median aggregate earnings growth is positive (negative) during periods of increasing (decreasing) rates. Further, the picture naming speed level is more negative during periods of decreasing rates than during periods of increasing rates.

Figure B1. Catch trial 1, always shown as the sixth target item of the questionnaire.

Results of these experiments reveal the sucrose preference of control mice was not different from that of stressed mice (Fig. 4E; t(14) = 2.14, p = 0.13). These results suggest that chronic predator stress does not induce anhedonia, a hallmark symptom of human depression [49–51]. Neurotrophins were not altered in response to chronic predator stress. Stress in humans is often associated with altered neurotrophin signaling [52,53]. Furthermore, brain neurotrophin signaling is altered in response to stress and also in depression, one of the adverse behavioral phenotype of stress disorders [54–56]. Therefore, using western blot analysis we measured the levels of NGF and BDNF in the hippocampus. Mature/pro-BDNF ratio of stressed mice was not different from that of the controls (Fig. 5A; t(6) = 2.44, p = 0.22). Similarly mature/precursor ratios of NGF were also not different between the stressed and control groups (Fig. 5B; t(6) = 2.44, p = 0.05). These results suggest BDNF and NGF signaling was not affected by the chronic stress. In the following, we discuss how the home country audit market was modified by chronic predator stress.

Figure B2. Catch trial 2, always shown as the eleventh target item of the questionnaire.

The adults' response patterns were compared with those in a previously published study that used a very similar set of visual and audio stimuli. Kurumada et al. (2004) conducted a laboratory-based eye-tracking experiment using the current construction "it looks like an X!" and "it LOOKS like an X..." while, unlike the current study, providing no feedback after each trial. They found that adult participants selected the mentioned animal on 65% of critical trials with noun-focus prosody, but only 13% of trials with verb-focus prosody. Adult participants in the current experiment selected the mentioned animal on 90% of the two practice trials and on 83% of the fourteen critical trials with noun-focus prosody, and on 78% of the practice trials and 59% of the critical trials with verb-focus prosody (Figure 5). Taken together, these results support the view that constant feedback helps adult listeners derive clearly distinguished interpretations based on noun-focus versus verb-focus intonation contours. Children, however, do not seem to use the feedback as effectively as adults do. We conducted a heat shock gene expression with the response data for critical trials produced by the four-year-olds (Gelman & Hill,1998).

Figure B3. Catch trial 3, always shown as the fifteenth target item of the questionnaire.

7 General Discussion

Contents

7.1 RQ1: the evidence related to CNC processing 221
7.2 RQ2: the evidence related to CNC use 223

As I described in the introduction, this thesis has two goals. The first goal (Goal 1), associated with Research Questions 1 (RQ1) and 2 (RQ2) is to better understand a structure I refer to as *complex nominal compounds* (CNC), which I described in detail in Chapter 2. Despite their common use in scientific texts, little is known yet about their online processing or about their distribution in these texts. Knowing more about this structure might contribute to improving the scientific register, popularly known to be hard to read.

In addition, my second goal (Goal 2) is to test two (different, but strongly related) previously proposed frameworks of language processing and use: the Entropy Rate Constancy (ERC) Principle posed by Genzel and Charniak (2002), and the Uniform Information Density (UID) Hypothesis posed by Jaeger (2010), which I described in Sections 3.3 and 3.4, respectively. In a nutshell, these frameworks presume that human communication is efficient, that speakers avoid transmitting too much information at once (i.e., avoid "peaks" of information density), and that, when they do, communication disruptions might occur. Since CNCs, as argued in Section 3.2.2 "condense large amounts of information in few words" (Jullian, 2001, p. 239), probably constituting "peaks" in information density, they are a good tool for testing predictions that can be derived from these frameworks.

The studies reported in this thesis put together these two goals by investigating the research questions (of Goal 1) while deriving predictions from the frameworks I wanted to test (Goal 2). In the following, I briefly summarize the results of these studies and how they answer the research questions. I organize the chapter based on the questions, but, since the studies' predictions and their results have implications for the validity of the to-be-tested frameworks, these implications are also included along in the research question discussions.

Again, for the sake of convenience, I restate each research question at the beginning of each section.

7.1 RQ1: the evidence related to CNC processing

RQ1 How are CNCs *processed* during reading? In particular, do complex nominal compounds pose difficulties for L1 and L2 sentence processing, as suggested by the literature?

Studies 1 and 2, reported in Chapter 5, used eye-tracking to investigate the processing of CNCs during reading. As far as I am aware, this is the first time the difficulties associated with CNC reading have been investigated with an online measure either in an L1 or in an L2. We compared the CNCs with another structure referred to as NPP (noun followed by prepositional phrase), and derived predictions stemming from the UID Hypothesis applied to comprehension. This aimed additionally at filling a gap in the literature, since the UID Hypothesis has not been investigated very much from the point of view of comprehension. The predictions were the following:

1. Readers should have longer reading times in the areas of text following the CNCs (the difficulty is predicted to "spill over" to the subsequent text areas);
2. Readers should make more regressions towards the CNCs;
3. Longer structures should also lead to longer reading times and more regressions;
4. For the L2 speakers, we should find an interaction between group and structure, indicating that L2 speakers whose L1 does not use CNCs often have more difficulty with them than L2 speakers whose L1 does make frequent use of CNCs.

The overall pattern of results found in both papers, however, was not very clear. For prediction 1, while we did find that CNCs led to longer reading times than NPPs, this was the case in only one reading measure in one experiment with L1 speakers (Study 1 was composed of two experiments), and only one reading measure in one experiment with L2 speakers (Study 2 also was composed of the same two experiments, but performed with L2 speakers), and these were not the same measures in the different groups, and happened in different experiments.

For prediction 2, concerning regressions, our findings were actually the reverse of what had been predicted: we consistently found (across all experiments in both Studies 1 and 2) more regressions towards NPPs than towards CNCs. In this case, however, we believe this was due to differences in the position of the head noun in the two structures. That is, while the head noun of CNCs is typically at the end of the CNC, the head noun of NPPs is typically at their beginning. Since the critical structure (i.e., the CNC or the NPP) was typically followed by a tense (i.e., *is / are*), readers probably needed to regress towards the beginning of the NPP in order to "revisit" the head noun and parse the agreement between the head noun and the tense; but this was not needed for CNCs because their head noun had been just read.

For prediction 3, we did find a clear and consistent effect of length on the number of regressions towards the critical region, but no effect of length on the other reading measures, related to reading times.

For prediction 4, we did not find the predicted interaction in either experiment of Study 2. Indeed, surprisingly, in one experiment, we found that the Portuguese/Spanish group regressed generally more than the German group towards the critical region (regardless of whether the sentence contained a CNC or an NPP), but that the German group spent longer times in the critical region (also regardless of the critical structure). While this may reflect some different strategies in the way that different L2 speaker groups may process these structures, we hesitate to interpret these effects, since we had no reason to predict them, and we only found them in one experiment.

What do these results mean for the UID Hypothesis? In Study 1 we argued that these results constituted at best weak support for the UID Hypothesis. This relatively weak support, however, may have been caused by the fact that comprehension (in this case, reading) is much less cognitively demanding than production. This was, in part, the reasoning behind Study 2. Study 2 was conceived as an attempt to have participants perform the exact same tasks, but in circumstances that would lead the task to be perceived as "more difficult", i.e., performing the task in a second language. However, Study 2 also yielded results that only weakly supported the UID Hypothesis.

This leads us to two possible interpretations. One possibility is that reading in an L2 is still not difficult enough to cause difficulty that is detectable through the eye-tracking measures we used. Maybe the UID Hypothesis does hold for comprehension, but we need a task with higher cognitive load requirements. For example, we could maybe ask participants to keep certain words in memory while reading, to see if this does lead to more clearly observable differences between the processing of CNCs and the processing of NPPs. Alternatively, maybe the UID Hypothesis simply does not hold as clearly for comprehension as it does for production. In this sense, further research is needed to better understand these results.

Overall, and responding to the research question more directly, it looks like CNCs are processed not very differently from other structures, such as NPPs, but do cause slightly more difficulty than NPPs, at least in the kinds of sentences used in the experiments reported in Studies 1 and 2, where the critical structure (CNC or NPP) appeared at the beginning of the sentence, without any supporting context. However, we found no evidence that this difficulty is modulated by the way CNCs are used in speakers' L1. In this sense, we did find some evidence in favor of the UID Hypothesis, but this evidence was not very strong, and it is still unclear how well it holds for comprehension. Further research is necessary to better understand why this evidence was not strong, and how the limitations of Studies 1 and 2 (such as the position of the head noun) impacted these results.

7.2 RQ2: the evidence related to CNC use

RQ2 How are CNCs *used* in scientific articles? In particular, how are they distributed through scientific articles, how are they set up, and how does their set up influence the difficulty perceived by readers when understanding them?

As mentioned earlier, despite the fact that CNCs are frequently used in scientific papers, little is known about the way they are distributed. Aiming to fill this gap, in Study 3, reported in Chapter 6, we set out to examine some characteristics of this distribution. For that, we constructed a corpus of academic papers containing papers from the fields of Biology, Economics and Linguistics and identified the CNCs present in the corpus. Based on the ERC Principle (and, of course, the assumption that CNCs are informationally dense), the following three predictions were posed:

1. CNCs should increase in frequency from the beginning to the middle to end of scientific papers;
2. CNCs should be set up by their context;
3. Once a CNC is used, it should repeat frequently;

Study 3 aimed at answering the first part of RQ2, related to how CNCs are *distributed* in scientific texts. The second part, related to how the support given by the context to a CNC influences the difficulty perceived by readers when encountering them, is answered by the two experiments reported in Study 4. In these experiments, participants read CNCs in the context where they originated and responded as to how difficult they found it to understand the CNCs. The CNCs used in the experiments were selected from the corpus constructed in Study 3. The contexts surrounding the CNCs (i.e., the actual text in the scientific papers from which the CNCs were selected) were categorized into different groups, according to how much support they provided to the CNC. This categorization was performed using a measure of contextual support calculated using word vectors produced by a large language model.

Study 4's Experiment 2 differs from Experiment 1 in two important ways, that will be relevant in the discussion below. First, it additionally considers the role of familiarity, that had been neglected in Experiment 1. CNCs were additionally categorized on whether they were familiar or not familiar, the measure of familiarity being based on Google Ngram Viewer (Michel et al., 2011) and the presence of the CNC or of its parts on Wikipedia. Second, in addition to responding how difficult they found it to understand the CNC, participants were also asked to produce paraphrases that explained how the words of the CNC were connected. The paraphrases were then tagged by three annotators as to whether they were correct or not, and the analysis only considered paraphrases that were deemed correct. This guaranteed that participants did not find "easy" CNCs that they had actually interpreted incorrectly.

Based on the UID Hypothesis, we thus posed our last prediction:

4. Readers should experience less difficulty when reading CNCs that received a lot of contextual support than when reading CNCs that received little contextual support.

For prediction 1, we found no significant increase in the numbers of CNCs, either between the beginning (the Introduction) and the middle of the papers (Method and Results), or between the middle and the end of the papers (Discussion and Conclusion). This was consistent across the different academic fields investigated. It is unclear why this was the case.

One possibility (suggested by a reviewer in an early submission of the paper) is that this might reflect the way in which academic papers are structured, and how the different sections of the papers typically have different roles that require the paper authors to write in very different ways. For example, in a typical scientific paper, the Introduction poses a research question and explains how the research question is related to other studies (that have already been published), as well as how this question has not yet been answered by them. This differs considerably from the Methods and Results sections (here taken as the middle part of the papers), which are much more direct (commonly following a boilerplate structure, say, describing participants, procedure, analysis, etc.), and often contain only a small amount of additional information, typically in the form of citations (which assume that the interested reader will be willing to read the cited papers if they want to know more) about the apparatus used in the experiment and the types of statistics performed to analyze the data. Thus, it is not surprising that the Methods and Results sections do not necessarily build much upon the Introduction, and do not have more CNCs than the Introduction. However, this would not explain why the end of the papers (comprising the Discussion and Conclusion sections) do not have more CNCs: even if the Discussion and Conclusion do contain some boilerplate structure, and even if the Discussion often reintroduces certain topics from the Introduction that might have already been forgotten by the reader, the Discussion normally does contain a substantial amount of new content that certainly builds upon the previous sections, and thus should have been found to have more CNCs than those previous sections.

A similar possibility, discussed in Study 3, is that the different parts of an academic paper might focus on different audiences. For example, the Introduction would be written for a broader audience, and the Methods would be written for more specialized readers. The problem with this idea, however, is the same as the one mentioned in the previous paragraph: it does not explain why the Conclusion does not have more CNCs than the previous sections.

Finally, a third possibility, that might only partially explain the negative result of prediction 1, is that certain CNCs (especially those that appear often in a paper) might become acronyms. In that case, even if there were new CNCs added to the text in the middle and in the end of the papers, the final number of CNCs would not be affected much. However, as we will see in prediction 3, not many CNCs repeat often. Therefore, even if the frequently repeated CNCs did become acronyms, the impact of creating these acronyms would be minimal on the final number of CNCs in a given section of a paper. Further research is necessary to understand why the number of CNCs does not seem to increase towards the end of scientific papers.

For prediction 2, our qualitative analysis of the way CNCs were used in the corpus papers showed that CNCs are indeed often introduced by their contexts (around 64% of the CNCs analyzed followed what we termed a "Gradual presetting", in which the preceding text slowly built up to the point where the CNC was clearly understandable). Most of the remaining CNCs (that were not introduced) either did not seem to require context support (e.g., *specific brain areas*), or used field-specific terminology (presumably assumed to be understood by the reader, e.g., *mammalian suprachiasmatic nucleus*), or used general scientific terminology (e.g., *baseline comparison condition*).

For prediction 3, CNCs were not found to repeat often. The vast majority (96.5%) of the CNCs in the corpus appeared up to three times in a given scientific paper. This goes against the ideas discussed in Section 2.6, which suggested that one of the functions of CNCs is to create new terms that can be subsequently referred to. Again, it is not clear why this is the case.

One possibility is that, instead of being repeated in full form, CNCs undergo what Norman (2003) termed **reductive head-repetition**. That is, a CNC such as *waste water treatment facility*, instead of reappearing with all of its words, may be just as easily (in contexts where there is no ambiguity) repeated as a *treatment facility*, or even just as *facility*. The discussion of Section 2.6 might actually provide the reason for this kind of abbreviation: if one of the reasons for using CNCs is to save space, skipping some of their words may be the obvious next step for saving even *more* space.

Turning to Study 4, we initially found (in Experiment 1) no effect of contextual support on the difficulty reported by participants. However, once we controlled for CNC familiarity (Experiment 2) and only considered responses that had been correctly paraphrased, we found that, while familiar CNCs were not affected by the contextual support, unfamiliar CNCs were.

What does this all mean for the ERC Principle? The results associated with prediction 2 do support the ERC Principle. CNCs do not come "out of the blue", and seem to follow a pattern that "smooths" the amount of information conveyed by them by accumulating into the global context enough content about their meaning before they are first used. However, the results associated with predictions 1 and 3 are at odds with the ERC Principle. While for prediction 3 the phenomenon of reductive head-repetition may explain our negative results, further research is needed to identify the reasons why the number of CNCs do not increase towards the end of scientific papers, as one would expect if the ERC Principle were true. Thus, taken as a whole, these results do partially support the ERC Principle, but raise questions about its full validity.

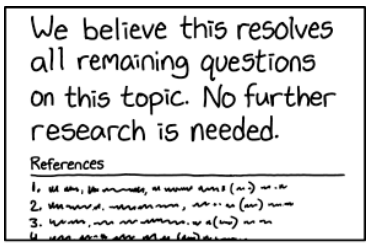
What about the UID Hypothesis? In addition to the discussion of the previous paragraph (that is also valid for the UID Hypothesis), the results of Study 4 do generally favor the UID Hypothesis, in its version related to comprehension.

Overall, CNCs are distributed roughly evenly through scientific articles, do not repeat frequently, and are mostly gradually introduced in the preceding text. This gradual set up has an effect on the perceived difficulty in comprehending unfamiliar CNCs but not on familiar ones. The results of Study 3 do favor the ERC Principle and the UID Hypothesis, but raise some questions about their validity. Those of Study 4 generally favor the UID Hypothesis.

8 Conclusion

Further Research is Needed

Mouseover text: Further research is needed to fully understand how we managed to do such a good job.



Munroe, 2020b

The work presented in this dissertation was motivated by the view, shared by many academics, that scientific texts are hard to read. The research was guided by the belief that it is probably worth understanding what the characteristics of these texts are if we are to find ways to improve this status quo. It was also guided by two models of language processing and use, and the predictions that they produced. My goals, therefore, were to further, even if just a little, the scientific understanding of one particular aspect of scientific texts (namely their frequent use of Complex Nominal Compounds), and to test the validity of these two models.

I thus posed two research questions, that accompanied us through the thesis. The first of them inquired about how CNCs are processed, or, more specifically, whether they are structures that cause processing difficulty to readers. Given the arguments of Chapter 2 (especially in Section 2.5) and Chapter 3 (especially Sections 3.2.2 and 3.4), CNCs were predicted to be difficult to process, and this was partially born out by the data (in Chapter 5), both for L1 and for L2 speakers.

The second question was related to how CNCs are used: how they are distributed, whether they repeat often, whether they are typically set up by their preceding context, and whether this context set up influences the difficulty readers perceive when reading them. Again, given the arguments in Chapter 2 (especially in Section 2.6) and Chapter 3 (especially Sections 3.2.2 and 3.3), CNCs were predicted to cluster towards the end of papers, to repeat often, to be set up by the context, and for the context to influence reading difficulty. While the first two of these predictions were not supported by the studies (in Chapter 6), the latter two were.

The results of this thesis offer an empirical validation of some recommendations typically found in writing guides. These guides often suggest that too long compounds should be avoided, or used with parsimony not to overwhelm readers. While these results do indicate that CNCs are hard to process, they also suggest that this difficulty can be mitigated by the support offered by the context when encountering them, and that familiar CNCs may not need to be avoided as much as unfamiliar ones. It is my hope that future writing guides will take these findings into consideration.

The results of this thesis also offer an empirical validation of the two models that were used to derive predictions for the studies reported in Chapters 5 and 6. They investigated the Uniform Information Density Hypothesis from the point of view of comprehension (a perspective that received little attention so far in the existing literature) and applied the Entropy Rate Constancy Principle to a domain that had not been considered yet, namely the use of CNCs. With these results, I hope to have contributed to the body of research investigating the application of information theoretic models of language processing and use in the field of Psycholinguistics.

The data presented in this thesis, however, are just a small step towards our understanding of these structures, and of scientific text in general. Nominal compounds have been studied for a long time (many of the papers I referred to in this thesis are many decades old), and the publications in this thesis do not settle any important scientific debate. Indeed, they actually raise *more* questions: why did the data seem to not support so strongly the UID Hypothesis? Why did we not find a cross-linguistic influence of speakers' L1 on the L2 processing of CNCs? Why did CNCs not repeat often, as had been predicted? Why do CNCs not cluster towards the end of scientific papers? Was this the result of the way we identified the CNCs in the corpus? Or could these unpredicted results be explained by a lack of power? I leave these questions for further research(ers) to consider, accepting that every project needs to eventually come to an end, and that this is the end of this one.

Bibliography

Algeo, J., & Algeo, A. S. (Eds.). (1991). *Fifty years among the new words: A dictionary of neologisms 1941-1991.* Cambridge: Cambridge University Press.

Alley, M. (1996). *The craft of scientific writing.* New York, NY: Springer.

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (Sixth ed.). Washington, DC: American Psychological Association.

Antoniová, V. K. (2020). An onomasiological approach to nominal compound semantics. *Word Structure*, *13*(3), 316–346. doi: https://doi.org/10.3366/word.2020.0174

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31–56. doi: https://doi.org/10.1177/00238309040470010201

Bailey, S. (2011). *Academic writing: A handbook for international students* (Third ed.). New York: Routledge.

Baldwin, T., & Tanaka, T. (2004). Translation by machine of complex nominals: Getting it right. In T. Tanaka, A. Villavicencio, F. Bond, & A. Korhonen (Eds.), *Proceedings of the Workshop on Multiword Expressions: Integrating Processing* (pp. 24–31). Stroudburg, PA: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/W04-0404`

Barnett, A., & Doubleday, Z. (2020). The growth of acronyms in the scientific literature. *Elife*, *9*, e60080. doi: https://doi.org/10.7554/eLife.60080

Bartolic, L. (1978). Nominal compounds in technical English. In L. Trimble, M. Trimble, & K. Drobnic (Eds.), *English for specific purposes: Science and technology* (pp. 257–277). Corvallis, OR: Oregon State University.

Bauer, L. (2003). *Introducing Linguistic Morphology.* Edinburgh: Edinburgh University Press. doi: https://doi.org/10.1515/9781474464284

Bauer, L., Lieber, R., & Plag, I. (2013). *The Oxford reference guide to English morphology.* Oxford: Oxford University Press.

Bauer, L., & Tarasova, E. (2013). The meaning link in nominal compounds. *SKASE Journal of Theoretical Linguistics*, *10*(3), 2–18. Retrieved from `http://www.skase.sk/Volumes/JTL24/pdf_doc/01.pdf`

Bell, M. J. (2012). The English noun-noun construct: a morphological and syntactic object. In A. Ralli, G. Booij, S. Scalise, & A. Karasimos (Eds.), *Proceeding of the Eighth Mediterranean Morphology Meeting (MMM8)* (Vol. 8, pp. 59–91). Cagliari: University of Patras. Retrieved from `https://pasithee.library.upatras.gr/mmm/article/view/2424/2683`

Benjamin, S., & Schmidtke, D. (2023). Conceptual combination during novel and existing compound word reading in context: A self-paced reading study. *Memory & Cognition*, *51*, 1170–1197. doi: https://doi.org/10.3758/s13421-022-01378-z

Berg, T. (2012). The cohesiveness of English and German compounds. *The Mental Lexicon*, *7*(1), 1–33. doi: https://doi.org/10.1075/ml.7.1.01ber

Berg, T. (2016). The semantic structure of English and German compounds: Same or different? *Studia Neophilologica*, *88*(2), 148–164. doi: https://doi.org/10.1080/00393274.2015.1135758

Bhatia, V. K. (1992). Pragmatics of the use of nominals in academic and professional genres. In L. F. Bouton & Y. Kachru (Eds.), *Pragmatics and language learning: Monograph*

series (Vol. 3, pp. 217–230). Urbana, Illinois, USA: University of Illinois. Retrieved from https://eric.ed.gov/?id=ED395531

Biber, D., & Gray, B. (2011). Grammatical change in the noun phrase: The influence of written language use. *English Language & Linguistics*, *15*(2), 223–250. doi: https://doi.org/10.1017/S1360674311000025

Brekle, H. E. (1976). *Generative Satzsemantik im System der englischen Nominalkomposition* [Generative sentential semantics in the English nominal system]. Munich: Fink.

Capurro, R., & Hjørland, B. (2003). The concept of information. *Annual Review of Information Science and Technology*, *37*(1), 343-411. doi: https://doi.org/10.1002/aris.1440370109

Carrió Pastor, M. L. (2008). English complex noun phrase interpretation by Spanish learners. *Revista Española de Lingüística Aplicada*, *21*, 27–44. Retrieved from https://dialnet.unirioja.es/servlet/articulo?codigo=2925910

Carrió Pastor, M. L., & Candel Mora, M. Á. (2013). Variation in the translation patterns of English complex noun phrases into Spanish in a specific domain. *Languages in Contrast*, *13*(1), 28–45. doi: https://doi.org/10.1075/lic.13.1.02car

Carroll, L. (1872). *Through the looking-glass.* London: Macmillan and Co. Retrieved from https://en.wikisource.org/wiki/Through_the_Looking-Glass,_and_What_Alice_Found_There

Clifton, C., Ferreira, F., Henderson, J. M., Inhoff, A. W., Liversedge, S. P., Reichle, E. D., & Schotter, E. R. (2016). Eye movements in reading and information processing: Keith Rayner's 40 year legacy. *Journal of Memory and Language*, *86*, 1–19. doi: https://doi.org/10.1016/j.jml.2015.07.004

Cohen, A. L., & Staub, A. (2014). Online Processing of Novel Noun–Noun Compounds: Eye Movement Evidence. *Quarterly Journal of Experimental Psychology*, *67*(1), 147–165. doi: https://doi.org/10.1080/17470218.2013.796398

Collins. (n.d.). Database. In *Collins english dictionary.* Retrieved January 4, 2024, from https://www.collinsdictionary.com/dictionary/english/database

Collins, M. X. (2014). Information density and dependency length as complementary cognitive models. *Journal of Psycholinguistic Research*, *43*, 651–681. doi: https://doi.org/10.1007/s10936-013-9273-3

Davies, M. (2004). British National Corpus. Oxford University Press. Retrieved from https://www.english-corpora.org/bnc/

de Almeida, R. G., Antal, C., & Salehi, K. (2025). Early morpho-orthographic and semantic effects in word recognition: Evidence from a foveal-splitting dichoptic paradigm with anaglyphs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. doi: https://doi.org/10.1037/xlm0001533

de Almeida, R. G., Dumassais, S., & Antal, C. (2020). Morphological parsing by foveal split: Evidence from anaglyphs. In S. Deninson, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 3055–3061). Cognitive Science Society. Retrieved from https://escholarship.org/uc/item/2f588344

Doubleday, Z. A., & Connell, S. D. (2017). Publishing with objective charisma: breaking science's paradox. *Trends in Ecology & Evolution*, *32*(11), 803–805. doi: https://doi.org/10.1016/j.tree.2017.06.011

Downing, P. (1977). On the creation and use of English compound nouns. *Language*, *53*(4), 810–842. doi: https://doi.org/10.2307/412913

Dressler, W. U. (2006). Compound types. In G. Libben & G. Jarema (Eds.), *The representation and processing of compounds words* (pp. 23–44). New York: Oxford. doi: https://doi.org/10.1093/acprof:oso/9780199228911.003.0002

Dubois, B. L. (1982). The construction of noun phrases in biomedical journal articles. In J. Hoedt, L. Lundquist, H. Picht, & J. Qvistgaard (Eds.), *Proceedings of the 3rd European Symposium on Language for Special Purposes: Pragmatics and LSP* (pp. 49–67). Copen-

hagen: School of Economics.

Fernández-Domínguez, J. (2010). N+N compounding in English: semantic categories and the weight of modifiers. *Brno Studies in English*, *36*(1), 47–76. Retrieved from `http://hdl.handle.net/11222.digilib/105088`

Ferčec, Y., Ivanka ; Liermann-Zeljak. (2015). Nominal compounds in technical English. In A. Akbarov (Ed.), *The Practice of Foreign Language Teaching: Theories and Applications* (pp. 268–277). Newcastle upon Tyne: Cambridge Scholars Publishing.

Floridi, L. (2009). Philosophical conceptions of information. In G. Sommaruga (Ed.), *Formal Theories of Information. Lecture Notes in Computer Science* (Vol. 5363, pp. 13–53). Heidelberg: Springer, Berlin. doi: https://doi.org/10.1007/978-3-642-00659-3_2

Frank, A. F., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the annual meeting of the Cognitive Science Society* (Vol. 30). Retrieved from `https://escholarship.org/uc/item/7d08h6j4`

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11. doi: https://doi.org/10.1016/j.bandl.2014.10.006

Fromkin, V., Rodman, R., & Hyams, N. (2011). *An Introduction to Language* (9e ed.). Boston: Wadsworth.

Gagné, C. L., & Shoben, E. J. (1997). Influence of thematic relations on the comprehension of modifier–noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(1), 71–87. doi: https://doi.org/10.1037/0278-7393.23.1.71

Gagné, C. L., & Spalding, T. L. (2009). Constituent integration during the processing of compound words: Does it involve the use of relational structures? *Journal of Memory and Language*, *60*(1), 20–35. doi: https://doi.org/10.1016/j.jml.2008.07.003

Gagné, C. L., & Spalding, T. L. (2006). Conceptual combination: implications for the mental lexicon. In G. Libben & G. Jarema (Eds.), *The representation and processing of compounds words* (pp. 125–144). New York: Oxford. doi: https://doi.org/10.1093/acprof:oso/9780199228911.003.0007

Gagné, C. L., & Spalding, T. L. (2013). Conceptual Composition: The Role of Relational Competition in the Comprehension of Modifier-Noun Phrases and Noun–Noun Compounds. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. 59, pp. 97–130). Amsterdam: Academic Press. doi: https://doi.org/10.1016/B978-0-12-407187-2.00003-4

Gallager, R. G. (1968). *Information theory and reliable communication*. New York: Wiley.

Gazni, A. (2011). Are the abstracts of high impact articles more readable? Investigating the evidence from top research institutions in the world. *Journal of Information Science*, *37*(3), 273–281. doi: https://doi.org/10.1177/0165551511401658

Geer, S. E., Gleitman, H., & Gleitman, L. (1972). Paraphrasing and remembering compound words. *Journal of Verbal Learning and Verbal Behavior*, *11*(3), 348–355. doi: https://doi.org/10.1016/S0022-5371(72)80097-5

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In P. Isabelle, E. Charniak, & D. Lin (Eds.), *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 199–206). Stroudsburg, PA: Association for Computational Linguistics. doi: https://doi.org/10.3115/1073083.1073117

Gerrig, R. J., & Bortfeld, H. (1999). Sense creation in and out of discourse contexts. *Journal of Memory and Language*, *41*(4), 457–468. doi: https://doi.org/10.1006/jmla.1999.2656

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056. doi: https://doi.org/10.1073/pnas.121643811

Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5),

389–407. doi: https://doi.org/10.1016/j.tics.2019.02.003

Glasman-Deal, H. (2010). *Science research writing for non-native speakers of English*. London: Imperial College Press.

Goldberg, A. E., & Shirtz, S. (2025). The English phrase-as-lemma construction: When a phrase masquerades as a word, people play along. *Language*, *101*(2), 291–320.

Granville Hatcher, A. (1960). An introduction to the analysis of English noun compounds. *Word*, *16*(3), 356-373. doi: https://doi.org/10.1080/00437956.1960.11659738

Günther, C., Kotowski, S., & Plag, I. (2020). Phrasal compounds can have adjectival heads: evidence from english. *English Language & Linguistics*, *24*(1), 75–95.

Hartley, J., Sotto, E., & Pennebaker, J. (2002). Style and substance in psychology: Are influential articles more readable than less influential ones? *Social Studies of Science*, *32*(2), 321–334. doi: https://doi.org/10.1177/0306312702032002005

Hillier, A., Kelly, R. P., & Klinger, T. (2016). Narrative style influences citation frequency in climate change science. *PLOS ONE*, *11*(12), 1-12. doi: https://doi.org/10.1371/journal.pone.0167983

Horsella, M., & Pérez, F. (1991). Nominal compounds in chemical English literature: Toward an approach to text typology. *English for Specific Purposes*, *10*(2), 125–138. doi: https://doi.org/10.1016/0889-4906(91)90005-H

Hrdina, C., & Hrdina, R. (2009). *Scientific English für Mediziner und Naturwissenschaftler* [Scientific English for doctors and scientists]. Berlin: Langenscheidt.

Isabelle, P. (1984). Another look at nominal compounds. In Y. Wilks (Ed.), *Proceedings of the 10th international conference on computational linguistics and 22nd annual meeting of the association for computational linguistics* (pp. 509–516). Stroudsburg, PA: Association for Computational Linguistics. doi: https://doi.org/10.3115/980491.980600

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62. doi: https://doi.org/10.1016/j.cogpsych.2010.02.002

Jarema, G. (2006). Compound representation and processing: A cross-language perspective. In G. Libben & G. Jarema (Eds.), *The representation and processing of compounds words* (pp. 1–22). New York: Oxford. doi: https://doi.org/10.1093/acprof:oso/9780199228911.003.0003

Jespersen, O. (1942). *A modern English grammar on historical principles*. Copenhagen: Ejnar Munksgaard. Retrieved September 18, 2025, from `https://archive.org/details/a-modern-english-grammar-on-historical-principles-part-vi/mode/2up`

Jullian, P. (2001). Mental representation of English complex nominals by Spanish speakers. *Onomázein*, *6*, 239–247. doi: https://doi.org/10.7764/onomazein.6.13

Juzek, T. S. (2024). Signal smoothing and syntactic choices: A critical reflection on the UID hypothesis. *Open Mind: Discoveries in Cognitive Science*, *8*, 217–234. doi: https://doi.org/10.1162/opmi_a_00125

Katz, M. J. (2009). *From research to manuscript: A guide to scientific writing* (Second ed.). Cleveland, OH: Springer.

Kerz, E. (2004). The cognitive and pragmatic motivations for the use of nominalizations in academic texts. In S. Burgess & P. Martín-Martín (Eds.), *English as an additional language in research publication and communication* (pp. 123–137). Bern, Switzerland: Peter Lang AG. Retrieved from `https://www.peterlang.com/document/1104303`

Kunter, G., & Plag, I. (2016). Morphological embedding and phonetic reduction: the case of triconstituent compounds. *Morphology*, *26*, 201–227. doi: https://doi.org/10.1007/s11525-016-9284-5

Kuperman, V., & Bertram, R. (2012). Moving spaces: Spelling alternation in English noun-noun compounds. *Language and Cognitive Processes*, *28*(7), 939–966. doi: https://doi.org/10.1080/01690965.2012.701757

Kvam, A. M. (1990). Three-part noun combinations in English, composition – meaning – stress. *English Studies: A Journal of English Language and Literature*, *71*(2), 152-161. doi: https://doi.org/10.1080/00138389008598684

Lashley, K. S. (1949). Persistent problems in the evolution of mind. *The Quarterly Review of Biology*, *24*(1), 28-42. doi: https://doi.org/10.1086/396806

Lees, R. B. (1960). *The grammar of English nominalizations.* Bloomington: Indiana University Press.

Levi, J. N. (1978). *The syntax and semantics of complex nominals.* New York: Academic Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. doi: https://doi.org/10.1016/j.cognition.2007.05.006

Levy, R., & Jaeger, T. F. (2006). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. C. Platt, & T. Hoffman (Eds.), *Proceedings of the 19th International Conference on Neural Information Processing Systems* (pp. 849–856). Cambridge, MA: MIT Press. Retrieved from `https://proceedings.neurips.cc/paper/2006/hash/c6a01432c8138d46ba39957a8250e027-Abstract.html`

Lewand, R. (2000). *Cryptological Mathematics.* The Mathematical Association of America.

Libben, G. (2006). Why study compound processing? An overview of the issues. In G. Libben & G. Jarema (Eds.), *The representation and processing of compounds words* (pp. 1–22). New York: Oxford. doi: https://doi.org/10.1093/acprof:oso/9780199228911.003.0001

Libben, G., Derwing, B. L., & de Almeida, R. G. (1999). Ambiguous novel compounds and models of morphological parsing. *Brain and Language*, *68*(1), 378-386. doi: https://doi.org/10.1006/brln.1999.2093

Libben, G., Gagné, C. L., & Dressler, W. U. (2020). The representation and processing of compounds words. In V. Pirrelli, I. Plag, & W. U. Dressler (Eds.), *Word knowledge and word usage: A cross-disciplinary guide to the mental lexicon* (pp. 336–352). Berlin, Boston: De Gruyter Mouton. doi: https://doi.org/10.1515/9783110440577-009

Lieber, R., & Štekauer, P. (2011). Introduction: Status and Definition of Compounding. In R. Lieber & P. Štekauer (Eds.), *The Oxford Handbook of Compounding.* Oxford University Press. doi: https://doi.org/10.1093/oxfordhb/9780199695720.013.0001

Limaye, M., & Pompian, R. (1991). Brevity versus clarity: The comprehensibility of nominal compounds in business and technical prose. *The Journal of Business Communication*, *28*(1), 7–21. doi: https://doi.org/10.1177/002194369102800102

Limerick, P. (1998). Dancing with professors: the trouble with academic prose. In V. Zamel & R. Spack (Eds.), *Negotiating Academic Literacies: Teaching and Learning across Languages and Cultures* (pp. 207–215). New York: Routledge.

Linh, N. M. (2010). *Noun-noun combinations in technical English.* (Master's thesis, Suranaree University of Technology, Nakhon Ratchasima, Thailand). Retrieved from `https://sutir.sut.ac.th/jspui/handle/123456789/3713`

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms.* Cambridge: Cambridge University Press.

Marchand, H. (1955). Notes on nominal compounds in present-day English. *Word*, *11*(2), 216–227. doi: https://doi.org/10.1080/00437956.1955.11659558

Marchand, H. (1960). *The categories and types of present-day English word formation.* Wiesbaden, Germany: Otto Harrassowitz. Retrieved from `https://archive.org/details/in.ernet.dli.2015.529076`

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330. Retrieved from `https://aclanthology.org/J93-2004/`

Marelli, M., Crepaldi, D., & Luzzatti, C. (2009). Head position and the mental representation of nominal compounds: A constituent priming study in Italian. *The Mental Lexicon*, *4*(3), 430–454. doi: https://doi.org/10.1075/ml.4.3.05mar

Maurits, L. (2012). *Representation, information theory and basic word order* (Doctoral dissertation, University of Adelaide, Adelaide, Australia). Retrieved from https://hdl.handle.net/2440/74128

Meister, C., Pimentel, T., Haller, P., Jäger, L., Cotterell, R., & Levy, R. (2021). Revisiting the Uniform Information Density hypothesis. In M.-F. Moens, X. Huang, L. Specia, & S. W. tau Yih (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 963–980). Stroudsburg, PA: Association for Computational Linguistics. doi: http://doi.org/10.18653/v1/2021.emnlp-main.74

Merriam-Webster. (n.d.-a). Database. In *Merriam-webster.com dictionary.* Retrieved January 4, 2024, from https://www.merriam-webster.com/dictionary/database

Merriam-Webster. (n.d.-b). Tl;dr. In *Merriam-webster.com dictionary.* Retrieved August 22, 2025, from https://www.merriam-webster.com/dictionary/TL%3BDR

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, T. G. B., ... Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176–182. doi: https://doi.org/10.1126/science.1199644

Montero, B. (1996). Technical communication: Complex nominals used to express new concepts in scientific English-causes and ambiguity in meaning. *The ESPecialist*, *17*(1), 57–72. Retrieved from https://revistas.pucsp.br/index.php/esp/article/view/9476/7042

Montgomery, S. L. (2003). *The Chicago guide to communicating science.* Chicago: The University of Chicago Press.

Moon, R. (1997). Vocabulary connections: multi-word items in English. In *Vocabulary: Description, acquisition and pedagogy* (pp. 40–63). Cambridge: Cambridge University Press.

Moroschan, G., Nicoladis, E., & Anjomshoae, F. (2024). Do children treat adjectives and nouns differently as modifiers in prenominal position? *Journal of Child Language*, 1–15. doi: https://doi.org/10.1017/S0305000924000448

Munroe, R. (2009). *The Sun.* XKCD, *673*. Retrieved November 16, 2025, from https://xkcd.com/673/

Munroe, R. (2018). *Boathouses and Houseboats.* XKCD, *2043*. Retrieved November 16, 2025, from https://xkcd.com/2043/

Munroe, R. (2020a). *Five Word Jargon.* XKCD, *2326*. Retrieved November 16, 2025, from https://xkcd.com/2326/

Munroe, R. (2020b). *Further Research is Needed.* XKCD, *2268*. Retrieved November 16, 2025, from https://xkcd.com/2268/

Munroe, R. (2023). *Garden Path Sentence.* XKCD, *2793*. Retrieved August 21, 2025, from https://xkcd.com/2793/

Munroe, R. (2024). *1.2 Kilofives.* XKCD, *2946*. Retrieved November 16, 2025, from https://xkcd.com/2946/

Nagy T., I., & Vincze, V. (2013). English nominal compound detection with Wikipedia-based methods. In *Text, Speech, and Dialogue (TSD) 2013. Lecture Notes in Computer Science* (pp. 225–232). Berlin: Springer. doi: https://doi.org/10.1007/978-3-642-40585-3_29

Nakov, P. (2013). On the interpretation of noun compounds: Syntax, semantics, and entailment. *Natural Language Engineering*, *19*(3), 291–330. doi: https://doi.org/10.1017/S1351324913000065

Norman, G. J. (2003). Consistent naming in scientific writing: sound adivce or shibboleth? *English for Specific Purposes*, *22*(2), 113-130. doi: https://doi.org/10.1016/S0889-4906(02)00013-3

Olshtain, E. (1981). English nominal compounds and the ESL/EFL reader. In M. Hines & W. Rutherford (Eds.), *On TESOL '81: Selected papers from the fifteenth annual Conference of Teachers of English to Speakers of Other Languages* (pp. 153–168). Washington, DC: TESOL. Retrieved from https://eric.ed.gov/?id=ED223079

Oxford. (n.d.). Database. In *Oxford english dictionary.* Retrieved January 4, 2024, from `https://www.oed.com/dictionary/database_n`

Palmer, H. E. (1917). *The scientific study & teaching of languages.* London: Butler and Tanner.

Pérez Ruiz, L. (2006). Unravelling noun strings: Toward an approach to the description of complex noun phrases in technical writing. *ES: Revista de Filología Inglesa*, *27*(1), 163–174. Retrieved from `https://dialnet.unirioja.es/servlet/articulo?codigo=2210385`

Piera, C. (1995). On compounding in English and Spanish. In H. Campos & P. Kempchinsky (Eds.), *Evolution and Revolution in Linguistic Theory* (pp. 302–315). Washington, D.C.: Georgetown University Press.

Pinker, S. (2014). *The sense of style: The thinking person's guide to writing in the 21st century.* New York: Penguin Books.

Plavén-Sigray, P., Matheson, G. J., Schiffler, B. C., & Thompson, W. H. (2017). Research: The readability of scientific texts is decreasing over time. *eLife*, *6*, e27725. doi: 10.7554/eLife.27725

Powell, A. L., Nelson, A. J., Hindley, E., Davies, M., Aggleton, J. P., & Vann, S. D. (2017). The rat retrosplenial cortex as a link for frontal functions: A lesion analysis. *Behavioural Brain Research*, *335*, 88–102.

Pueyo, I. G. (1996). The construction of technicality in the field of plastics: A functional approach towards teaching technical terminology. *English for Specific Purposes*, *15*(4), 251–278. doi: https://doi.org/10.1016/S0889-4906(96)00011-7

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422. doi: https://doi.org/10.1037/0033-2909.124.3.372

Richman, S. (1969). The translation to Spanish of English nouns in juxtaposition. *Hispania*, *52*(3), 426–430. doi: https://doi.org/10.2307/337897

Salager, F. (1984). Compound nominal phrases in scientific-technical literature: Proportion and rationale. In A. K. Pugh & J. M. Ulijn (Eds.), *Reading for professional purposes: Studies in native and foreign languages* (pp. 136–145). London: Heinemann.

Schmidtke, D., Kuperman, V., Gagné, C. L., & Spalding, T. L. (2016). Competition between conceptual relations affects compound recognition: The role of entropy. *Psychonomic Bulletin & Review*, *23*(2), 556–570. doi: https://doi.org/10.3758/s13423-015-0926-0

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. doi: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Siepmann, D., Gallagher, J. D., Hannay, M., & Mackenzie, L. (2011). *Writing in English: A guide for advanced learners.* Tübingen: A. Francke Verlag.

Sikos, L., Greenberg, C., Drenhaus, H., & Crocker, M. W. (2017). Information density of encodings: The role of syntactic variation in comprehension. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 3168–3173).

Spalding, T. L., Gagné, C. L., Mullaly, A., & Ji, H. (2010). Relation-based interpretation of noun-noun phrases: A new theoretical approach. In S. Olsen (Ed.), *New impulses in word-formation* (pp. 283–315). Hamburg: Buske.

StatQuest with Josh Starmer. (2021). *Entropy (for data science) clearly explained!!!* [Video]. Youtube. Retrieved from `https://www.youtube.com/watch?v=YtebGVx-Fxw`

Stremersch, S., Verniers, I., & Verhoef, P. C. (2007). The quest for citations: Drivers of article impact. *Journal of Marketing*, *71*(3), 171–193. doi: https://doi.org/10.1509/jmkg.71.3.171

Stricker, J., Chasiotis, A., Kerwer, M., & Günther, A. (2020). Scientific abstracts and plain language summaries in psychology: A comparison based on readability indices. *PLOS ONE*, *15*(4), 1-9. doi: https://doi.org/10.1371/journal.pone.0231160

Sword, H. (2012). *Stylish academic writing*. Cambridge, MA: Harvard University Press.

Tobin, M. J. (2002). Compliance (COMmunicate PLease wIth Less Abbreviations, Noun Clusters, and Exclusiveness). *American Journal of Respiratory and Critical Care Medicine*, *166*(12), 1534–1536. doi: https://doi.org/10.1164/rccm.2211001

van Helmond, K., & van Vugt, M. (1985). On the transferability of nominal compounds. *Interlanguage Studies Bulletin*, 5–34. Retrieved from `https://www.jstor.org/stable/43135308`

Vincze, V., Nagy T., I., & Berend, G. (2011). Multiword Expressions and Named Entities in the Wiki50 Corpus. In R. Mitkov & G. Angelova (Eds.), *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011* (pp. 289–295). Hissar, Bulgaria: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/R11-1040/`

Weiskopf, D. A. (2007). Compound nominals, context and compositionality. *Synthese*, *156*, 161–204. doi: https://doi.org/10.1007/s11229-005-3489-1

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & van den Bosch, A. (2016). Prediction during natural language comprehension. *Cerebral Cortex*, *26*(6), 2506–2516. doi: https://doi.org/10.1093/cercor/bhv075

Williams, R. (1984). A cognitive approach to English nominal compounds. In A. K. Pugh & J. M. Ulijn (Eds.), *Reading for professional purposes: Studies in native and foreign languages* (pp. 136–145). London: Heinemann.

Xu, Y., & Reitter, D. (2018). Information density converges in dialogue: Towards an information-theoretic model. *Cognition*, *170*, 147–163. doi: https://doi.org/10.1016/j.cognition.2017.09.018

Zimmer, K. E. (1971). Some general observations about nominal compounds. In *Working papers on language universals* (Vol. 5). Stanford University. Retrieved from `https://eric.ed.gov/?id=ED104124`

Curriculum Vitae

Personal Information

Name: John Cristian Borges Gambôa

Address: Erwin-Schrödinger-Str. 57–423 D-67663 Kaiserslautern

Email: gamboa@rptu.de

Education

2018 – 2026 Doctoral student, Faculty of Social Sciences
RPTU University of Kaiserslautern-Landau, Germany
Supervisor: Prof. Dr. Shanley Allen

2014 – 2017 Master of Science in Computer Science
University of Kaiserslautern (TUK), Germany

2008 – 2013 Bachelor in Computer Science
Federal University of Rio Grande do Sul (UFRGS), Brazil

Academic Employment

10.2021 – current Researcher (Wissenschaftliche Mitarbeiter)
Psycholinguistics and Language Development
University of Kaiserslautern-Landau (RPTU)

09.2014 – 06.2016 Research Assistant, Chair of Real Time Systems
University of Kaiserslautern, Germany

03.2015 – 09.2015 Research Assistant, Psycholinguistics and Language Development
University of Kaiserslautern, Germany

03.2013 – 12.2013 Teaching Assistant, Natural Language Processing Lab
Federal University of Rio Grande do Sul, Brazil

02.2012 – 01.2013 Research Assistant, Chair of Real Time Systems
University of Kaiserslautern, Germany

02.2011 – 12.2011 Research Assistant, Computer Graphics Research Group
Federal University of Rio Grande do Sul, Brazil

Publications

Gamboa, J. C. B., Fernandez, L. B., & Allen, S. E. M. (2025). *Investigating the Uniform Information Density Hypothesis in L2 with complex nominal compounds* [Manuscript submitted for publication]. Psycholinguistics and Language Development Research Group, RPTU University of Kaiserslauter-Landau.

Gamboa, J., Gvozdeva, D., Ito, F., Järvikivi, J. & Allen, S. E. M. (2025). *How does the context preceding a long nominal compound influence its comprehension difficulty?* [Manuscript submitted for publication]. Psycholinguistics and Language Development Research Group, RPTU University of Kaiserslauter-Landau.

Fernandez, L. B., Nota, N., **Gamboa, J.**, Amos, R. M., Corps, R. E., Hadley, L. V. & Pickering, M. J. (2025). *Visual World Paradigm and Divergence Point Analysis with Bootstrapping: Researcher Degrees of Freedom in Real and Simulated Data with Recommendations* [Manuscript submitted for publication]. Psycholinguistics and Language Development Research Group, RPTU University of Kaiserslauter-Landau.

Fernandez, L. B., Hadley, L. V., **Gamboa, J. C. B.**, Allison, C., & Allen, S. E. M. (2025). The impact of speech rate on first- and second-stage prediction in L1 and L2 speakers. *Bilingualism: Language and Cognition,* Advance online publication, 1–13. https://doi.org/10.1017/S1366728925100515

Gamboa, J. C. B., Fernandez, L. B., & Allen, S. E. M. (2024). Investigating the Uniform Information Density hypothesis with complex nominal compounds. *Applied Psycholinguistics, 45*(2), 322–367. https://doi.org/10.1017/S0142716424000092

Gamboa, J., Braun, K., Järvikivi, J. & Allen, S. E. M. (2024). The distributional properties of long nominal compounds in scientific articles: an investigation based on the uniform information density hypothesis. *Corpus Linguistics and Linguistic Theory, 21*(1), 137–171. https://doi.org/10.1515/cllt-2023-0028

Fernandez, L. B., Hadley, L. V., Koç, A., **Gamboa, J. C.**, & Allen, S. E. (2024). Is there a cost when predictions are not met? A VWP study investigating L1 and L2 speakers. *Quarterly Journal of Experimental Psychology, 0*(0). https://doi.org/10.1177/17470218241270200

Dash, A., Sahu, A., Shringi, R., **Gamboa, J. C. B.**, Afzal, M. Z., Malik, M. I., Dengel, A. & Ahmed, S. (2017). AirScript - creating documents in air. In *14th ICDAR International Conference on Document Analysis and Recognition* (pp. 908-913). https://doi.org/10.1109/ICDAR.2017.153

Binsfeld, R., **Gamboa, J.**, & Walter, M. (2011). Visual patterns in the plant kingdom. In *24th SIBGRAPI Conference on Graphics, Patterns and Images* (pp. 86-92). IEEE. https://doi.org/10.1109/SIBGRAPI.2011.44