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.

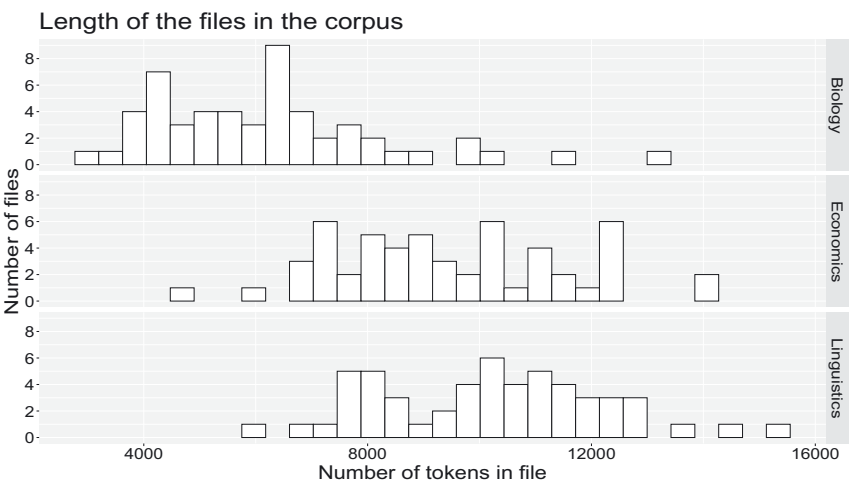


Figure 1: Histogram of the length of the corpus files.

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.

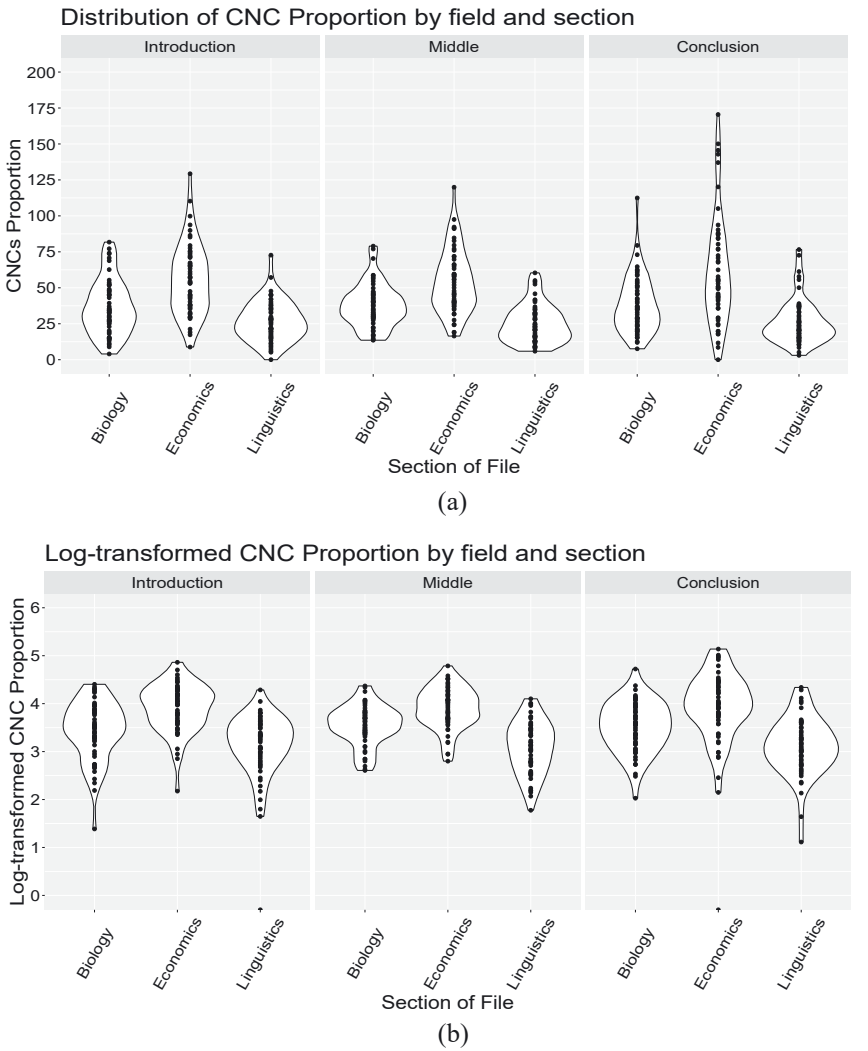


Figure 2: The CNC proportion by field and section. Each datapoint is a file section in the corpus. (a) Raw values; (b) log-transformed values.

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	***

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

$$\mathrm{DV} \sim \mathrm{section} + \mathrm{field} + (1 \mid \mathrm{file})$$

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.

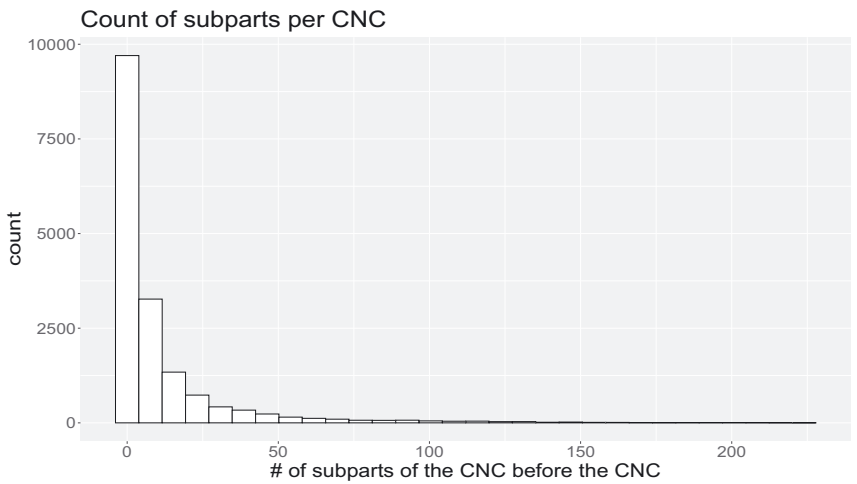
3.3 Are CNCs preceded by their subparts?

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.

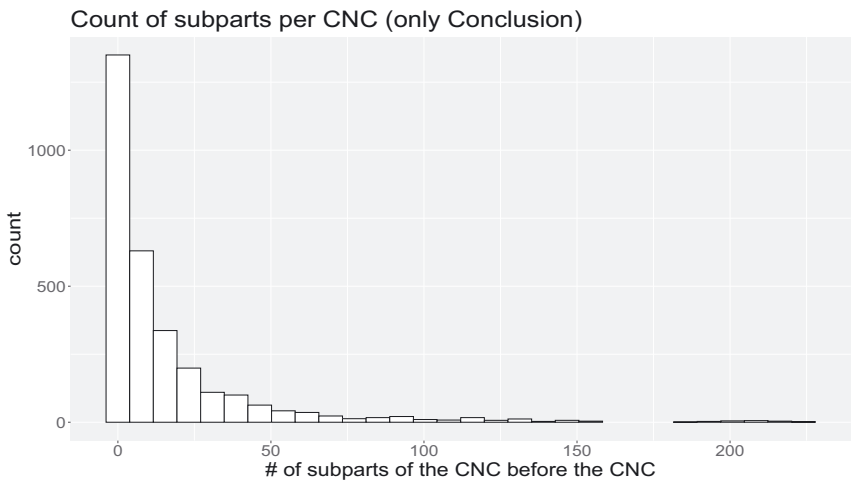
For each CNC, we count the number of bigram subparts that appeared in the whole text before the CNC. For example, given the CNC *left feeder reward probability*, we counted how many times *left feeder*, *feeder reward* and *reward probability* appeared in the whole portion of the article preceding it.[10] 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.



(a)



(b)

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

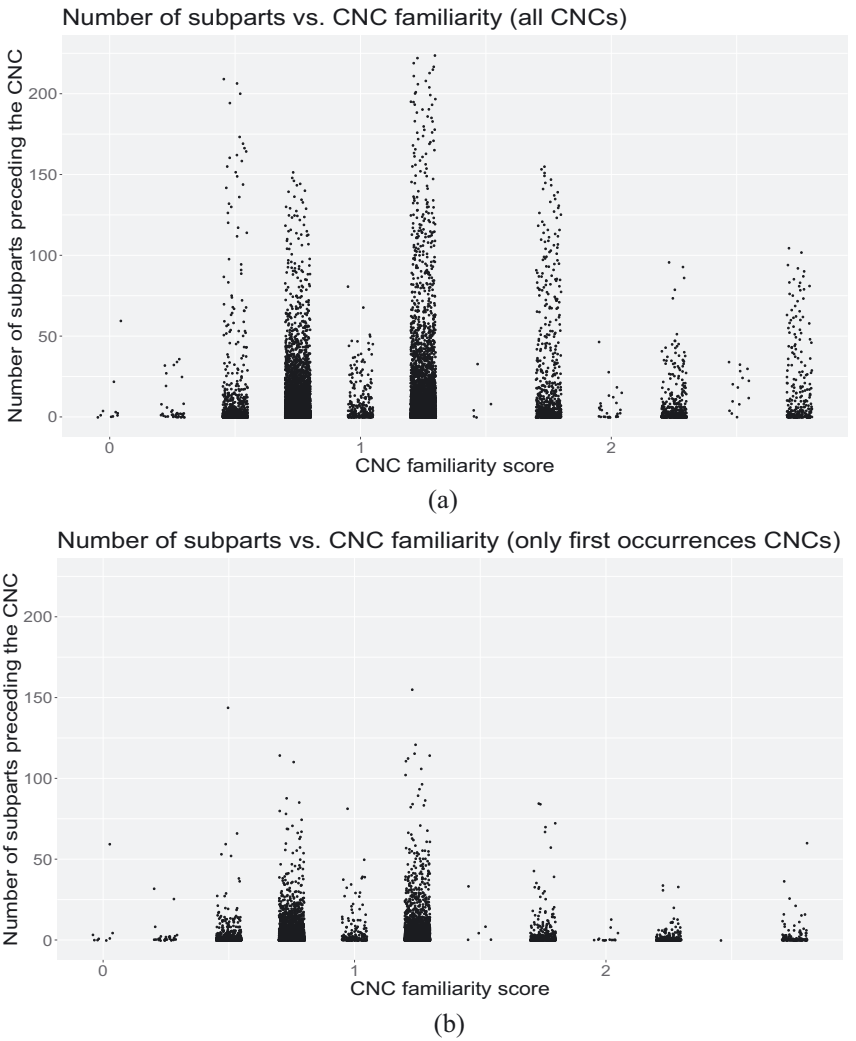


Figure 4: Points represent CNCs with a given familiarity score and a given number of preceding subparts. While in (a) we consider all 3-word CNCs (14,357 data points), in (b) we consider only the first occurrence of a CNC in a given article (9,956 data points).

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).