




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

© The Author(s), 2024. Published by Cambridge University Press

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,

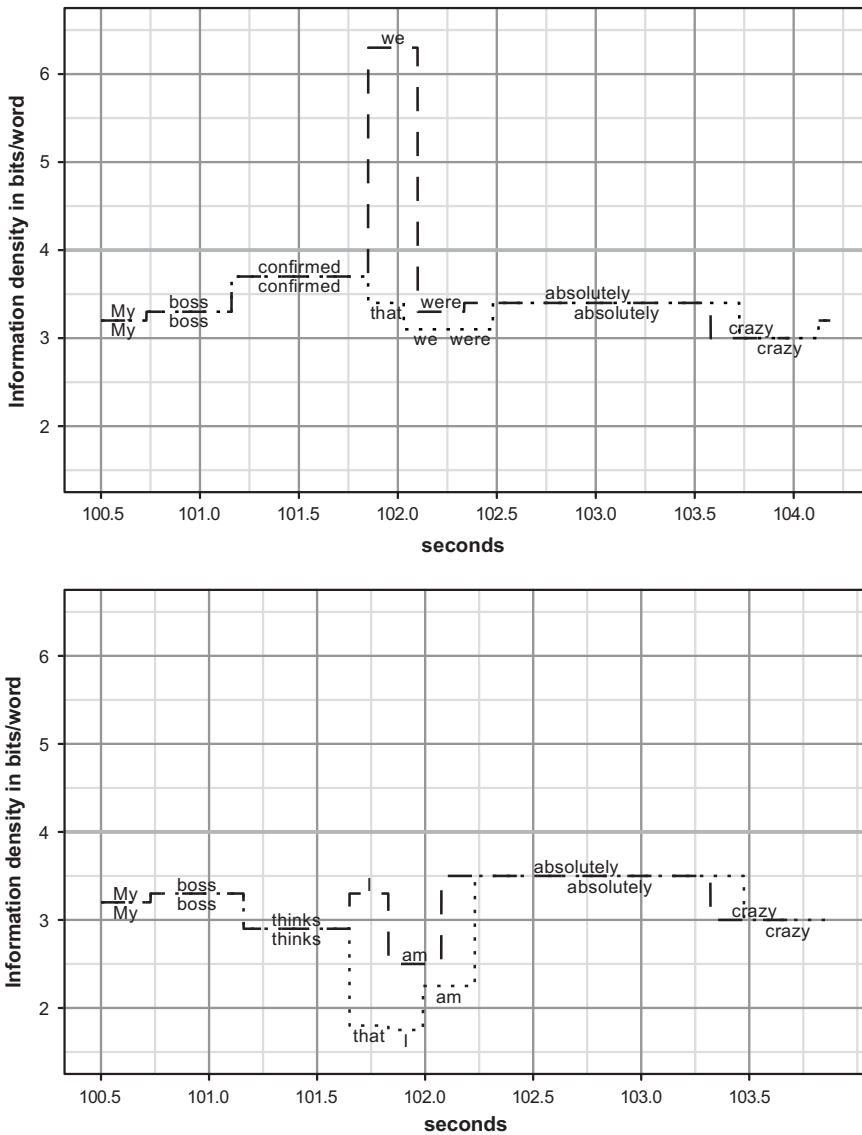


Figure 1. An example of the predictions of the UID hypothesis. Speakers have been shown to prefer using *that* in (a), but not in (b). Reprinted from Jaeger (2010, figure 1), with permission from Elsevier.

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

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

NCs¹ 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

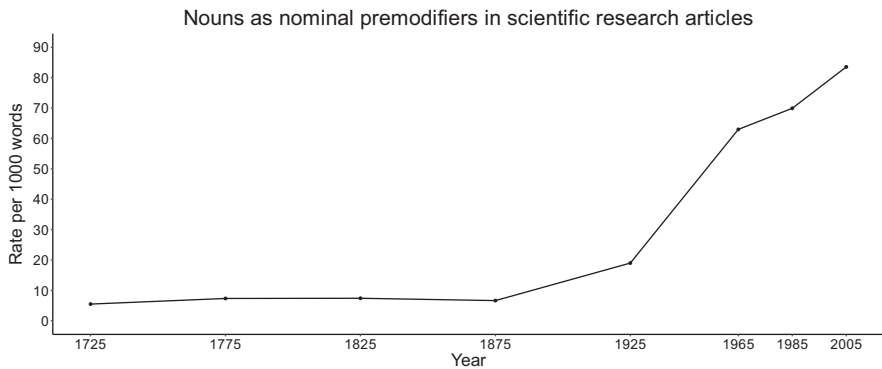


Figure 2. The development of the use of nouns as nominal premodifiers (i.e., forming NCs) through time. Adapted from (Biber & Gray, 2011).

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 (Brekke, 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

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; Olshstein, 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 study²

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

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

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				

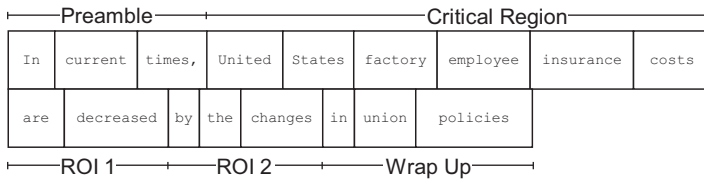


Figure 3. An example trial reconstructed from the data collected in Experiment 1. (The reconstruction's font size does not reflect the exact font size of stimulus presentation – note the varying distance between words.) The boxes around each word indicate the interest area associated with that word: 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](#)).

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

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).

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

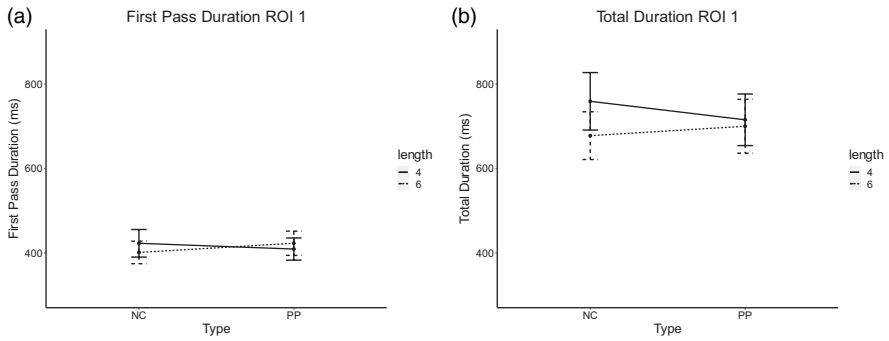


Figure 4. Experiment 1: Reading measures in Region of Interest 1. Error bars represent 95% confidence intervals.

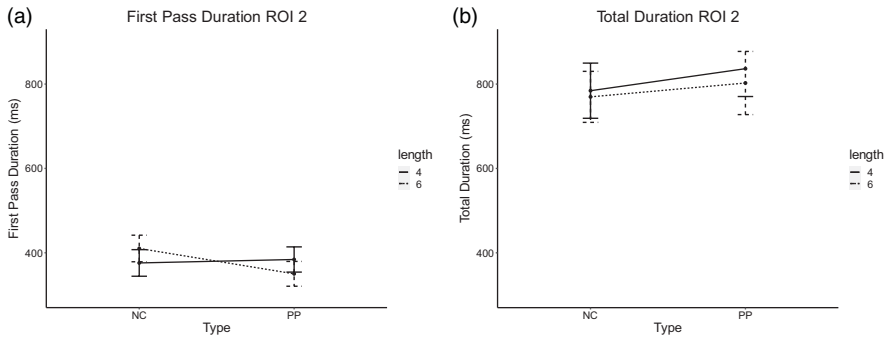


Figure 5. Experiment 1: Reading measures in Region of Interest 2. Error bars represent 95% confidence intervals.

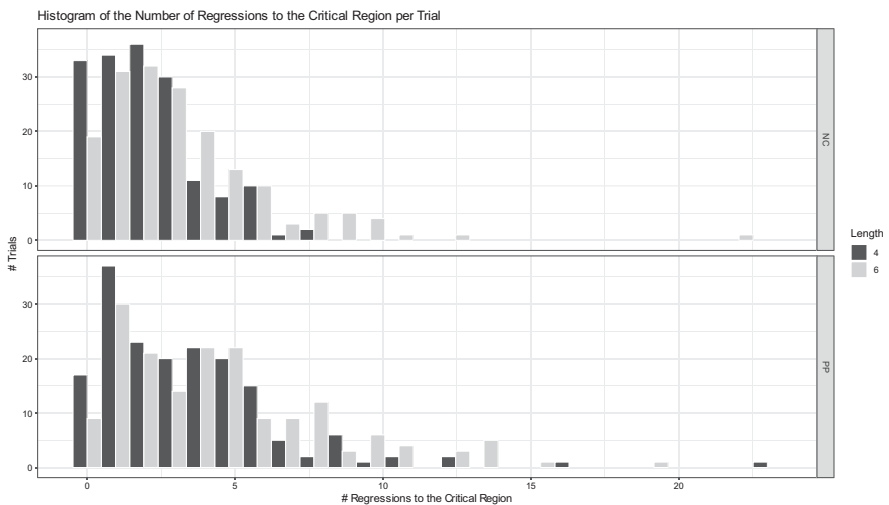


Figure 6. Experiment 1: Regressions onto the critical region.

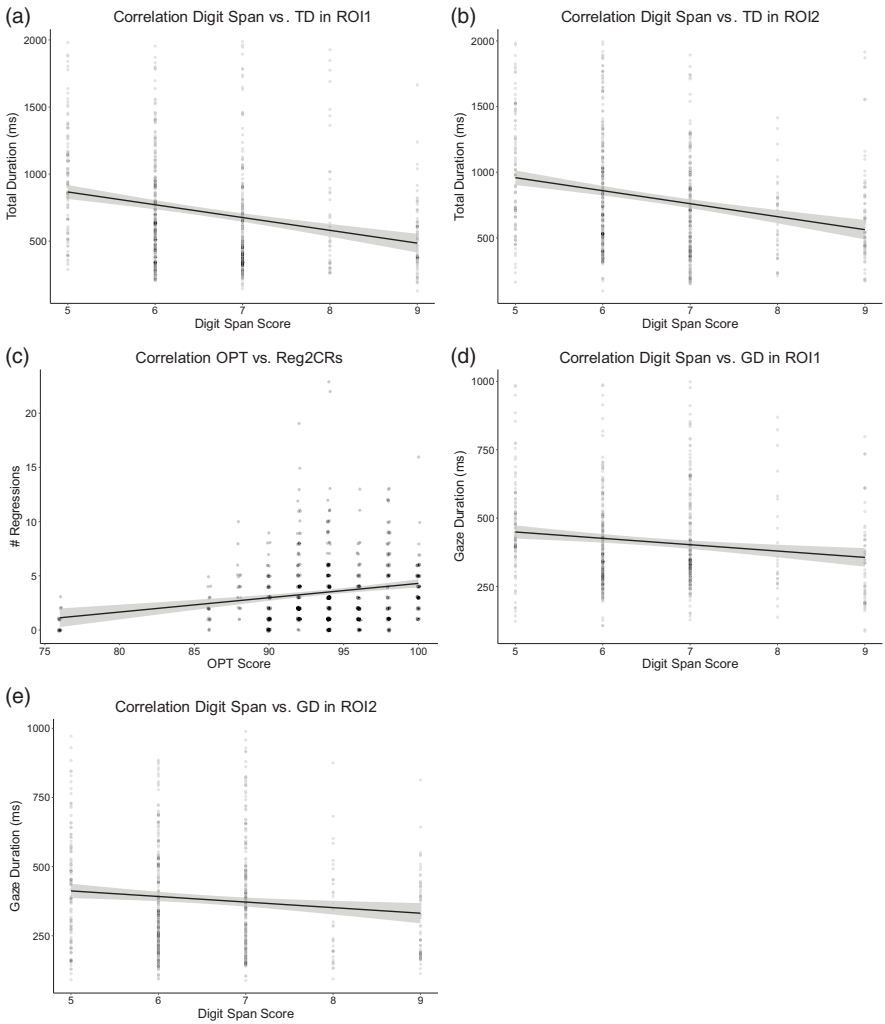


Figure 7. Experiment 1: The relation between the reading measures and some of the covariate data in our study. We use transparent data points in order to indicate their density. The covariates in (a), (b), and (c) were found to be significant predictors of the reading measures they are presented with. 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

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.

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 collocations⁶ (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.

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

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

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.

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 Factor⁸ (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

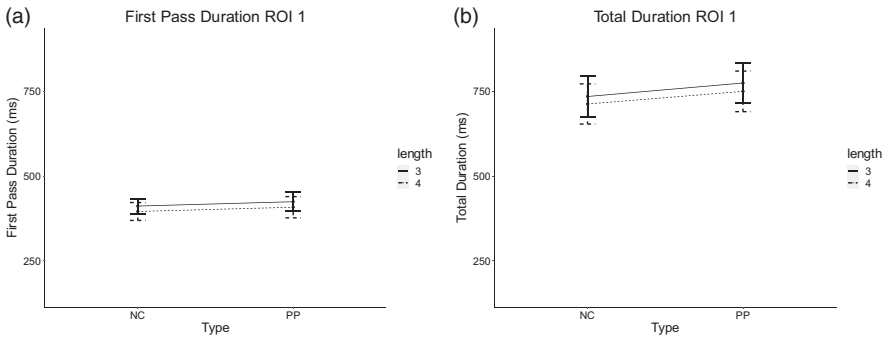


Figure 8. Experiment 2: Reading measures in Region of Interest 1. Error bars represent 95% confidence intervals.

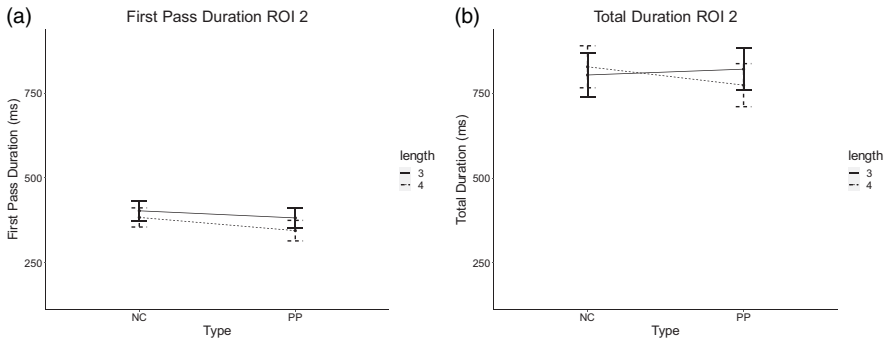


Figure 9. Experiment 2: Reading measures in Region of Interest 2. Error bars represent 95% confidence intervals.

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).

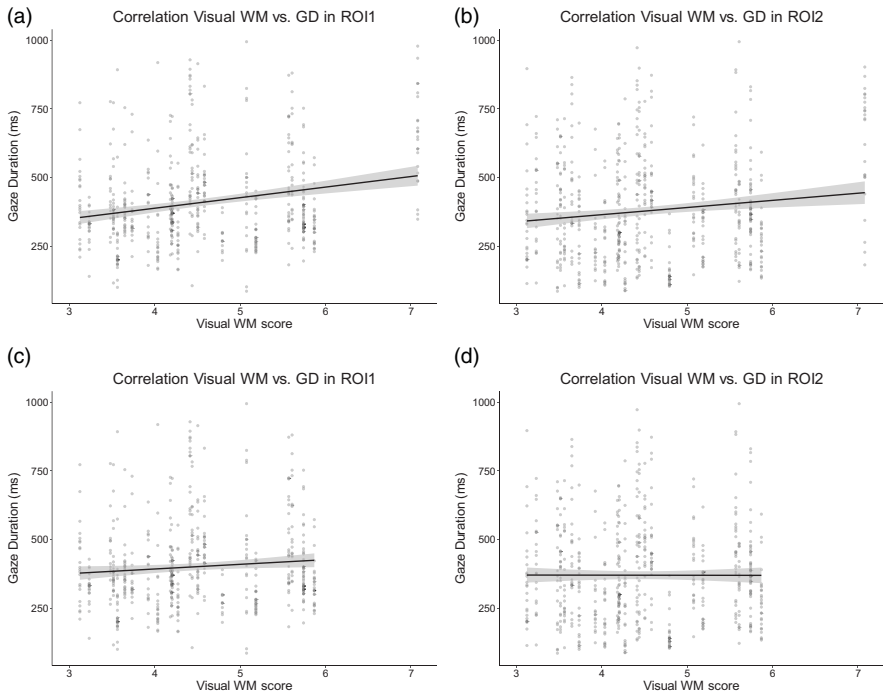


Figure 10. Experiment 2: Correlation between participants’ visual WM scores and First Pass Duration in ROI1 (a) and ROI2 (b), found to be significant predictors in our model, and the same results (c and d, respectively) after removing the single outlier with very high Visual WM score. Gray shading indicates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

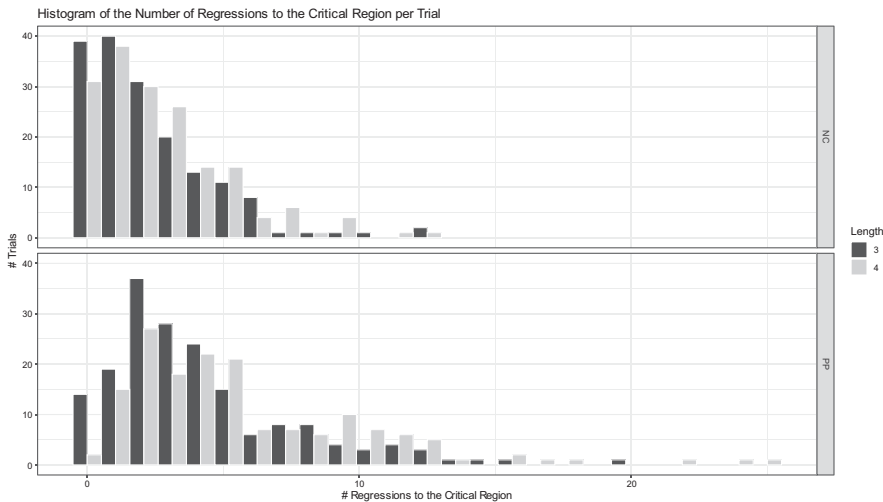


Figure 11. Experiment 2: Regressions onto the critical region.

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

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.

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

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

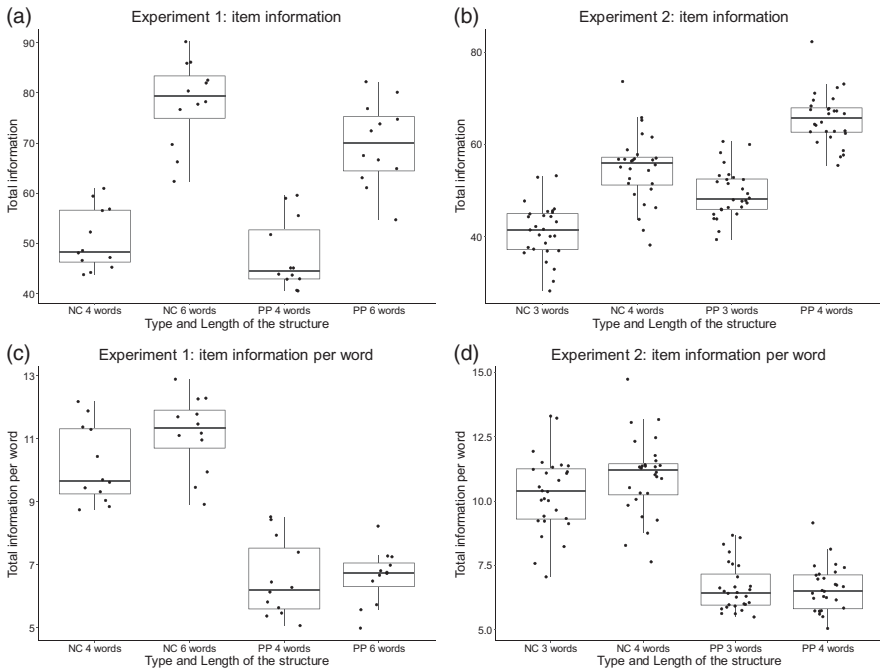


Figure 12. The information of items in Experiment 1 and 2. Graphs show the total amount (a and b) and the information per word (c and d). In all graphs, each point represents an item.

(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

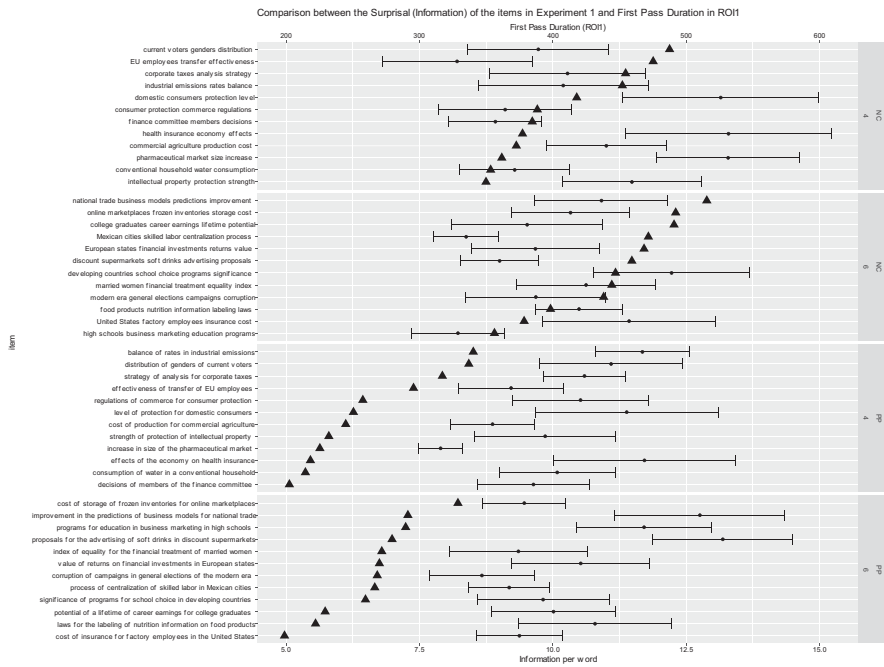


Figure 13. Figure shows items of Experiment 1 ordered by their surprisal per word, as indicated by the triangles, and separated by condition. Mean and standard error of FPD in ROI1 are indicated by the dot and the error bars. Similar comparisons between reading measures and *total* surprisal of each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

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 Utgaliyev, 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 $DV \sim OPT + DST + length * type + (1 + length * type | subject) + (1 + DST + OPT + type * length | item)$.

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 $DV \sim OPT + SpellingScore + length * type + VerbalWM + VisualWM + (1 + length * type | subject) + (1 + VerbalWM + VisualWM + type * length | item)$

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 & Surlis, 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>

- 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
- Brekke, 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., & Surlis, 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](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](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. B. (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/vjyfy>
- 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)

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)

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)

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)

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)

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

	<i>b</i>	SE	<i>t</i>	<i>p</i>
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

	<i>b</i>	SE	<i>t</i>	<i>p</i>
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

	<i>b</i>	SE	<i>t</i>	<i>p</i>
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

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

	<i>b</i>	SE	<i>t</i>	<i>p</i>
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

	<i>b</i>	SE	<i>z</i>	<i>p</i>
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

	<i>b</i>	SE	<i>z</i>	<i>p</i>
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

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

	<i>b</i>	SE	<i>t</i>	<i>p</i>
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

	<i>b</i>	SE	<i>t</i>	<i>p</i>
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

	<i>b</i>	SE	<i>t</i>	<i>p</i>
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

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

	<i>b</i>	SE	<i>t</i>	<i>p</i>
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

	<i>b</i>	SE	<i>z</i>	<i>p</i>
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)

	<i>b</i>	SE	<i>z</i>	<i>p</i>
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

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 $P(\text{"mat"} | \text{"the cat sat on the"}) = 0.0000000145$ (a very low probability), and the highest probability is assigned to the word "floor" ($P(\text{"floor"} | \text{"the cat sat on the"}) = 0.06707$). To calculate the information in "mat," we use

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

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, \dots, 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

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

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:

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

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>