# The inverse frequency effect

## An exploratory study

David Temperley
University of Rochester

Rare syntactic constructions show an especially strong tendency to be repeated, but some rare constructions exhibit this tendency much more strongly than others. The reasons for this variation are not well understood. This exploratory study examines five rare noun-phrase (NP) expansions in English: <*the* A> (*the rich*), <*a* $N_{prop}$ $N_{prop}$> (*a Bob Gates)*, <$N_{sing}$ $N_{prop}$ $N_{prop}$> (*architect Julia Morgan*), <D $N_{pl}$ $N_{sing}$> (*the jobs data*), and <$N_{sing}$ A $N_{sing}$> (*home electronic equipment*). Repetition tendencies are very strong in the first and second of these and somewhat strong in the third; in the fourth and fifth they are much weaker, only slightly higher than those of common NP expansions such as <D A $N_{sing}$> (*the black dog*). To explain this variation, we suggest that constructions may be associated with different types of discourse: constructions with high repetition tendencies tend to occur in persuasive rather than informative discourse.

**Keywords:** priming, discourse, coordination, parallelism, inverse frequency effect

## 1. Introduction

Syntactic constructions are often used in a repetitive manner: when a construction is used in discourse, this increases its probability of being used again in the near future. However, syntactic repetition is more likely to occur in some situations than in others. Several factors affecting syntactic repetition have been identified. Coordinate structures show an especially strong tendency for repetition, and rare syntactic constructions show a stronger repetition tendency than common ones; syntactic repetition can also be used for rhetorical purposes. In this study, we present a corpus analysis of NP expansions in written English that explores the role of all three of these factors in syntactic repetition. While the effects of coordinate parallelism and construction frequency have been explored

https://doi.org/10.1075/ijcl.22080.tem | Published online: 9 October 2023
*International Journal of Corpus Linguistics* ISSN 1384-6655 | E-ISSN 1569-9811

in prior studies, a novel contribution of the current study is to examine the role of discourse type: we find that the constructions showing the strongest repetition tendencies are those associated with rhetorical or 'persuasive' discourse.

## 2. Previous work on syntactic repetition

Much of the research on syntactic repetition has invoked the concept of 'priming' – the general idea that exposure to a stimulus (an object, event, or abstract structure) facilitates the perception and processing of further occurrences of that stimulus. One line of research in this area has focused on syntactic alternations: cases where there are two semantically equivalent (or nearly equivalent) ways of expressing the same thought. An example is the English dative construction, which can be realized by either a double-object form (*She gave me the book*) or a prepositional form (*She gave the book to me*): the contextual use of one form or the other encourages the subsequent use of that form (Bock, 1986). Priming effects have been found with a variety of syntactic alternations, both in experiments (Ferreira, 2003; Cleland & Pickering, 2003; Konopka & Bock, 2009) and in corpus data (Gries, 2005; Szmrecsanyi, 2005; Reitter et al., 2011; Jaeger & Snider, 2013).

Another line of research has focused on the repetition of syntactic structures more generally and the factors that influence it. One finding is that repetition is especially common in coordinate structures, a phenomenon sometimes known as 'parallelism'. In a sentence of the form NP *and* NP, it is highly probable that the second conjunct will syntactically match the first, e.g. *the brown dog and the black cat* (Dubey et al., 2008; Temperley & Gildea, 2015). Another finding is that rare syntactic constructions show especially strong priming effects, a phenomenon known as the 'inverse frequency effect'. In a study of dialogue, Reitter et al. (2006) found that low-frequency constructions show a stronger tendency than common ones to be repeated within a short time. And in a study of NP coordinate structures in newspaper text, Temperley and Gildea (2015) showed that the increased likelihood of an NP expansion occurring in the second conjunct when it occurred in the first conjunct is especially great for low-frequency expansions. Research on syntactic alternations has found evidence for the inverse frequency effect as well, although such work tends to focus on the relative frequency of two forms in an alternation (e.g. the double-object dative versus prepositional dative) rather than on their absolute frequency: the rarer of the two forms tends to prime more strongly (Hartsuiker et al., 1999; Ferreira, 2003; Scheepers, 2003; Jaeger & Snider, 2013).

Syntactic priming has been explained in a variety of ways. Some accounts have viewed it as arising from low-level, domain-general mechanisms such as activation (Pickering & Branigan, 1998; Reitter et al., 2011) or implicit learning (Chang et al., 2006; Jaeger & Snider, 2013). Other accounts have suggested that syntactic priming may serve communicative functions specific to language. Pickering and Garrod (2004) view priming as an aspect of 'alignment' – the tendency for participants in a dialogue to coordinate their linguistic representations at multiple levels. While this explanation does not apply in such an obvious way to written text, it may be that writers assume that their use of a construction will cause alignment on the reader's part, facilitating comprehension of subsequent uses of the construction. Along similar lines, Ferreira (2019) argues that syntactic priming can be viewed, at least in part, as a strategy to promote successful communication. For example, when one repeats a syntactic construction used by another speaker in conversation, it shows that the construction was heard and understood.

Another factor that may affect priming is 'uniform information density': the preference to maintain a moderate, even flow of linguistic information (in the probabilistic sense) (Fenk & Fenk, 1980; Levy & Jaeger, 2007). It is well-established that comprehension is affected by the frequencies and contextual probabilities of words and syntactic constructions; both unexpected words and rare constructions are processed more slowly (Trueswell & Tanenhaus, 1994; Hale, 2001; Levy, 2008). Production processes are sensitive to this, adjusting to extend high-information segments in time and compress low-information ones (Aylett & Turk, 2004; Frank & Jaeger, 2008). Temperley and Gildea (2015) invoke information flow to explain the inverse frequency effect in coordinate structures. There is a strong expectation for the syntactic structure of the second conjunct to match the first (Frazier et al., 2000); when a low-frequency construction is used in this way, it softens the "spike" in information that the construction might otherwise cause. While the repetition only directly reduces the information of the second conjunct, not the first, Temperley and Gildea argue that the processing of the first conjunct may "spill over" to the second, and that an information spike in the first conjunct may therefore be mitigated by low information in the second. Temperley and Gildea also examine two other predictions that follow from the information flow perspective: (i) that lexical probabilities will be lower in matching second conjuncts than in first conjuncts (since the low lexical probabilities of the second conjunct are counterbalanced by its high syntactic probability), and (ii) that lexical probabilities will be lower in matching second conjuncts than in non-matching ones. Both predictions are confirmed.

A final issue relevant to syntactic repetition is its use in rhetoric – as a means of persuasion, of showing emotional commitment on the part of the speaker/

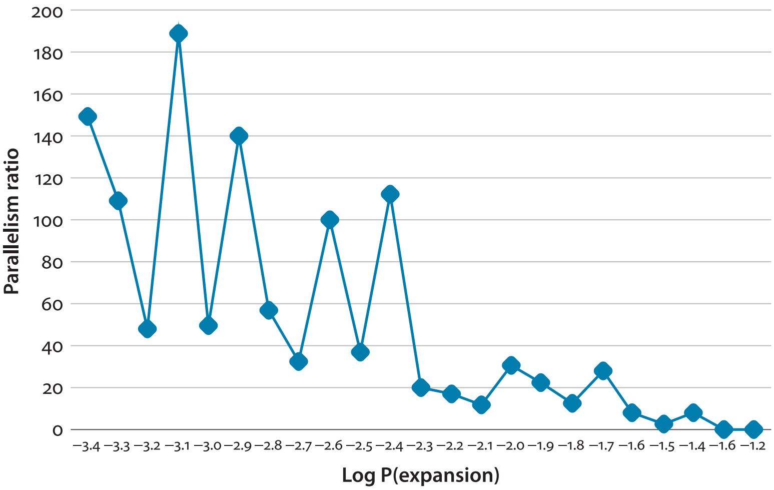
writer, or of arousing emotion in the listener/reader. The repetition of words, phrases, and syntactic structures has long been recognized as a powerful rhetorical device (Corbett, 1971; Vickers, 1994). McQuarrie and Mick (1996) observe many uses of repetition in advertising, including syntactic repetition, e.g. *The quality you need. The price you want.* In an experimental study, Menninghaus et al. (2017) find that removing devices of 'parallelistic diction' from poems – including syntactic repetition as well as rhyme, alliteration, meter, and other kinds of patterning – makes them less aesthetically appealing and emotionally impactful. On the other hand, it has been argued that another important aspect of rhetoric is 'artful deviation' – a stretching or violation of the usual rules, in syntax or other domains (Corbett, 1971; Vickers, 1994; McQuarrie & Mick, 1996). Thus, using a rare or even marginally grammatical syntactic construction in a repetitive fashion may be a natural way of combining two rhetorical devices – a potential explanation for the inverse frequency effect. Temperley (2022) suggests, based on informal observation, that many repetitive uses of rare constructions seem to be associated with what has been called 'persuasive' rather than 'informative' discourse (Brewer, 1980; Sanders, 1997).

In the current study, we offer a further exploration of the inverse frequency effect. We focus on the absolute frequency of constructions, rather than on the relative frequency of alternating forms (such as the double-object and prepositional forms of the dative construction). We address some limitations of prior work in this area and attempt to gain a deeper understanding of the factors influencing syntactic repetition. While the overall reality of the inverse frequency effect seems clear, it does not appear to be the only factor affecting the use of repetition in syntax. Figure 1 shows data from Temperley and Gildea's (2015) study, representing NP expansions in newspaper text. The horizontal axis represents the log probability of expansions, binned into categories with an increment of 0.1. (Figure 1 in Temperley & Gildea, 2015, represents the same data but with larger bins.) The vertical axis represents the 'parallelism ratio' of expansions:

parallelism ratio = $P(E \mid C_M) \,/\, P(E)$

where $E$ is an expansion and $C_M$ is a 'matching context' (one in which $E$ is the second conjunct of a coordinate phrase and the same expansion occurred in the first conjunct). For example, consider the NP expansion <D A $N_{sing}$> (e.g. *the big dog*). (D = determiner, A = adjective, $N_{sing}$ = singular noun; throughout the paper, expansions will be indicated in angle brackets.) This expansion has an overall probability (given an NP parent) of .031; its probability of occurring in a matching context is .069; so its parallelism ratio is .069 / .031 = 2.2. In Figure 1, the inverse frequency effect emerges very clearly; on the whole, the parallelism ratio strongly increases as expansion probability decreases. However, the effect is quite uneven;

even though each frequency bin represents multiple expansions, some bins have much higher parallelism ratios than others with similar probabilities. This suggests that the inverse frequency effect is much stronger for some rare expansions than others, and that the effect may be being driven by a fairly small number of expansions.



*Note.* For NP expansions in the *Wall Street Journal* corpus, this shows each expansion's 'parallelism ratio' – its probability of occurring in the second conjunct of a coordinate phrase after occurring in the first, divided by its overall probability – against the overall (log base 10) probability of the expansion, with expansions binned by frequency.

**Figure 1.** Data from Temperley & Gildea (2015)

In the current study we focus on a small set of NP expansions in English and examine their tendency to be repeated at short intervals – what we will call 'close repetition'. We consider both repetition in coordinate phrases and close repetition more generally. We predicted (following previous studies) that rare constructions would show stronger repetition tendencies than common ones, and we predicted that coordinate structures would account for a large proportion of these repetitions. We also expected there to be considerable variation among the rare constructions in these respects, and we hoped that this would shed light on the factors influencing syntactic repetition – factors that modulate or interact with the inverse frequency effect. Indeed, the rare constructions examined here vary widely in their tendency toward close repetition. We expected that this variation might be explained in rhetorical terms – that the expansions showing close repetition would tend to be those that are associated with persuasive rather than informative discourse; we find some support for this prediction. However, the identification of constructions as "persuasive" or "informative" is somewhat conjectural, subjective, and gradient; in light of this, it seemed questionable to use this as the basis for a hypothesis. Rather, it is a tentative explanation that requires further study. In that sense, the current study is best regarded as exploratory rather than hypothesis-driven. Apart from this, the current study provides a corpus of hand-checked tokens of rare syntactic constructions that may be useful for other purposes (available at http://davidtemperley.com/inverse-frequency; see Appendix B for details).

## 3. Methodology

In this section, we discuss how we chose which syntactic expansions to study and how we defined them. Then we explain how we measured close repetition and parallelism in each expansion.

### 3.1 Choice and definition of syntactic constructions

In exploring the repetition of syntactic constructions, how should syntactic construction types be defined? Previous corpus studies of NP expansions (Dubey et al., 2008; Temperley & Gildea, 2015) have defined constructions by the top-level expansion of the NP – its sequence of children. Similarly, Reitter et al. (2006), who examine repetition of syntactic "rules" more generally, define a rule as a parent constituent combined with its children. All three of these studies employ the widely used Penn Treebank corpus (Marcus et al., 1994) – either the *Wall Street Journal* portion of the corpus (hereafter the WSJ corpus), containing newspaper text (Dubey et al., 2008; Temperley & Gildea, 2015), or the *Switchboard* portion, containing telephone conversations (Reitter et al., 2006) – and assume the syntactic annotation system used there. This approach is problematic. For example, consider the NP expansion <NP SBAR>. In the Penn Treebank annotation system, SBAR is used for all finite dependent clauses; thus in the expansion <NP SBAR>, the SBAR might be a relative clause (Example 1a) (as assumed by Dubey et al.), but it could also be an infinitival clause (Example 1b) or a noun-complement clause (Example 1c) (examples are from the WSJ corpus):

(1) a. A soft landing is (NP (NP an economic slowdown) (SBAR that eases inflation without leading to a recession)).
b. Later yesterday, a Massachusetts senate committee approved ((NP a bill) (SBAR to allow national interstate banking by banks)) …
c. Don threw in the towel just about (NP (NP the time) (SBAR he should have doubled his bet)) …

One might question whether these are all instances of the same construction. We can simplify the situation somewhat by confining our attention to expansions that consist only of 'preterminals' (part-of-speech tags corresponding to single words) – sometimes known as "flat" or "base" NPs. Even then, there are complications. Flat NPs vary greatly in their number of children. In the WSJ corpus, there are 4020 different flat NP expansions; 56% of them occur only once. A large majority of the low-frequency expansions contain four or more children, such as those in Examples (2a) and (2b):

(2) a. <DT JJ JJ NNP NNP NN> (occurs twice): the notorious old Khmer Rouge leader
    b. <NN CC NN NN NN NNS> (occurs once): cogeneration and waste heat recovery plants

It seems problematic to compare expansions of very different lengths, since length itself may well be a factor influencing repetition. One might also say, in many cases, that long expansions have an internal hierarchical structure – e.g. *the notorious old* (*Khmer Rouge*) *leader*, ((*cogeneration*) *and* ((*waste heat*) *recovery*)) *plants* – but such structure is not represented in the Treebank annotations and is difficult to capture automatically.

An alternative to the "top-level expansion" approach is to define constructions in more intuitive terms, informed by linguistic theory and comprehensive taxonomic descriptions of English grammar such as Quirk et al. (1985) and Huddleston & Pullum (2002). But this, too, is not ideal. Frazier et al. (2000) found that, in an NP coordinate phrase in which the first conjunct's structure is <D A N>, processing of the second conjunct is facilitated when its structure is <D A N> versus <D N>. But the NP expansions <D A N> and <D N> are not well-established "natural kinds" in linguistic theory; there are no names for them, to our knowledge, nor for similar constructions such as <D A A N> or <D N N>. Thus, constructions that are not singled out in linguistic theory may still have cognitive importance. On the other hand, even something like <D A N> might subdivide into further constructions (depending, for example, on whether the determiner was definite or indefinite, or whether the noun was singular or plural); it is not known to what extent such distinctions affect repetition (in Frazier et al.'s study, the two conjuncts in a coordinate phrase always agreed in number and had the same determiner).

To our knowledge, there is no perfect solution to the problem of defining constructions. In the current study, we focus on NP expansions containing only preterminals – D (determiner), A (adjective), $N_{sing}$ (singular common noun), $N_{pl}$ (plural common noun), and $N_{prop}$ (proper noun) – with either two or three preterminals in each expansion. In so doing, we minimize length differences between constructions, and avoid constructions with complex hierarchical structure. For the most part, we define each construction in terms of its preterminals. However, certain constructions seem to be syntactically limited in the choice of determiners: one can say *the rich* but not *a rich*. In such cases, a specific determiner is part of the definition. Overall, then, the set of constructions used here (along with their definitions) is based on a combination of corpus annotation systems, conventional linguistic wisdom, and our own intuitions. Our choices might be criticized and are certainly not the only possibilities; in this regard, too, the current study may be viewed as exploratory.

As noted earlier, our focus is especially on rare NP expansions, though we also consider some common NP expansions for comparison. Perusal of natural written text suggests several NP expansions that seem uncommon or non-standard (again, this is based on intuition, though – as we will see – our intuitions are supported by corpus data). Examples are given in Table 1; natural instances of each one (from the WSJ corpus) are shown in the second column. In <$N_{sing}$ $N_{prop}$ $N_{prop}$>, a proper name is preceded by a common noun that clarifies its reference – sometimes known as a 'restrictive appositive'. In <*a* $N_{prop}$ $N_{prop}$>, a proper name carries an indefinite determiner. This rather playful construction recasts a singular definite referent as a category: the first example means *Someone like Bob Gates might even have said ....* (In both of these constructions, we assume that the proper name is two words; usually it refers to a person, though not always.) The expansion <*the* A> contains only the definite determiner followed by an adjective, sometimes known as a substantivized adjective. In <D $N_{pl}$ $N_{sing}$>, a noun carries a plural noun modifier (normally, noun modifiers are singular). And in <$N_{sing}$ A $N_{sing}$>, a noun modifier is followed by an attributive adjective (normally the adjective would precede the noun modifier). By Temperley and Gildea's (2015) argument, the rarity of these constructions (if they are indeed rare) should make them especially prone to repetition.

To explore these constructions systematically, we needed a way of identifying them in corpus data. The ideal corpus for such an investigation would be one showing both preterminals and constituent structure. However, corpora with hand-annotated constituent structure are relatively small: the largest, the WSJ corpus, has about 1.1 million words. For very rare syntactic constructions, there may not be enough tokens to draw conclusions about their use; our initial experiments with the WSJ corpus confirmed this. Larger corpora are available with automatically generated syntactic annotations (e.g. Charniak et al., 2000). But these annotations are not completely accurate, and given the probabilistic nature of modern parsing algorithms, we suspect that rare constructions – having rarely been seen in the algorithms' training – are more likely to be mislabeled.

**Table 1.** Rare NP expansions used in the study

| **Construction** | **Examples from WSJ corpus** |
|---|---|
| <$N_{sing}$ $N_{prop}$ $N_{prop}$> | *Cartoonist Garry Trudeau* is suing the Writers Guild of America ... |
| | Its fanciful offices were designed by *architect Julia Morgan* ... |
| <*a* $N_{prop}$ $N_{prop}$> | Why, *a Bob Gates* might even have said ... |
| | [He] seems to have had *a Jennifer Bartlett* in mind ... |
| <*the* A> | Its ambiguity and uneasy mixture of *the serious* and *the comic* is ... |
| | China penalizes *the efficient* and rewards *the incompetent* ... |
| <D $N_{pl}$ $N_{sing}$> | *The appropriations clause* states that ... |
| | The markets await tomorrow's release of *the jobs data* ... |
| <$N_{sing}$ A $N_{sing}$> | the federal government would provide $ 97 million in *emergency federal support* ... |
| | ... prices for *home electronic equipment* fell 1.1% ... |

To solve these problems, we used a combination of automatically annotated corpus data and hand-filtering. Our corpus is an 11-million-word sample of the Corpus of Contemporary American English (COCA, Davies, 2009). The COCA corpus (and our sample of it) contains eight different types of discourse in roughly equal amounts: academic, blog, fiction, magazine, news, spoken, TV/movies, and web. (The entire corpus contains over 1 billion words; hereafter, "the corpus" refers only to our 11-million-word sample.) The corpus is annotated with preterminals (automatically generated using the CLAWS-7 tagger, which yields 96–97% accuracy according to Garside & Smith, 1997) but not syntactic bracketing. We extracted tokens of rare NP expansions from the corpus by defining them as sequences of preterminals and/or specific words. We included some additional constraints to limit the number of false positives. For example, with the construction <*the* A>, we excluded tokens that were immediately followed by a noun, since we found that the target phrase in those cases was usually a modifier of the following noun (e.g. *the rich people*) rather than a stand-alone noun phrase. (While the pattern *the* A N could possibly represent a token of <*the* A> followed by an independent noun – as in the hypothetical sentence *We shouldn't give the rich benefits* – this seems very unlikely to occur; a search of the 1-million-word WSJ corpus did not find a single instance of this.) We then went through the tokens manually and removed any remaining false positives. Appendix A provides more detail about the search process we used for each of the rare expansions. Our publicly released data shows, for each rare expansion, all the tokens found by our automatic search processes as well as those that we judged to be valid tokens.

We also considered some common NP expansions for comparison with the rare ones. Here, we took the eight most common NP expansions in the WSJ corpus that (i) involved exactly two or three preterminals, and (ii) involved only the preterminal types used in the rare expansions: determiner, singular common noun, plural common noun, singular proper noun, and adjective. These types are shown in the first eight rows of Table 2. We also subdivided the construction <D A $N_{sing}$> into two more specific constructions – <*the* A $N_{sing}$> and <*a* A $N_{sing}$> – in order to have some constructions with specific determiners like the second and third rare constructions in Table 1. These are shown in the last two rows of Table 2. For these ten common NP expansions, manual inspection of all tokens in the COCA corpus was not possible due to the large numbers of tokens (the most common expansion, <D $N_{sing}$>, occurs over 300,000 times). However, inspection of samples suggested that automatic extraction of tokens produces very few false positives. Table 2 shows, for each common expansion, the number of false positives that were found in a hand-checked random sample of 100 tokens. (In distinguishing "true" from "false" tokens of an expansion, we follow the conventions used in the Penn Treebank; see Appendix A for details.)

**Table 2.** Common NP expansions used in this study

| **Construction** | **Examples from WSJ corpus** | **False positives*** |
|---|---|---|
| <D $N_{sing}$> | ... *a forum* likely to bring new attention to *the problem*. | 4 |
| <$N_{prop}$ $N_{prop}$> | *Pierre Vinken*, 61 years old, ... | 4 |
| <D A $N_{sing}$> | Despite *the gloomy forecast*, ... | 2 |
| <A $N_{pl}$> | ... *preliminary findings* were reported more than a year ago ... | 7 |
| <D $N_{pl}$> | 9.8 billion Kent cigarettes with *the filters* were sold ... | 1 |
| <A $N_{sing}$> | ... a forum likely to bring *new attention* to the problem. | 16 |
| <$N_{sing}$ $N_{pl}$> | ... a high percentage of *cancer deaths* ... | 14 |
| <D $N_{sing}$ $N_{sing}$> | South Korea has recorded *a trade surplus* ... | 4 |
| <*the* A $N_{sing}$> | Despite *the gloomy forecast*, ... | 8 |
| <*a* A $N_{sing}$> | ... *a high percentage* of cancer deaths ... | 1 |

* In a random set of 100 tokens extracted from the COCA corpus.

As noted earlier, many studies of syntactic repetition have focused on syntactic alternations – cases where two constructions are similar or identical in meaning, such as dative alternation. We do not claim that the constructions studied here are involved in syntactic alternations. There are some loose connections between rare and common constructions. For example, <*the* A> can sometimes be regarded as a substitution for either <*the* A $N_{sing}$> or <A $N_{pl}$>; however, this substitution can only be made under certain circumstances (discussed further in Appendix A). Thus, comparing one NP expansion to another with regard to repetition tendencies may be of limited interest. Our focus, rather, is on general patterns: whether the rare constructions show stronger repetition tendencies than the common ones, and whether the variation among the rare constructions in this regard points to any possible explanation.

## 3.2 Measuring close repetition and parallelism

In analyzing the data, we had two main aims. First, we wished to examine the tendency for each NP expansion to be used in close repetition – what we will call its 'repetition tendency'. We measured this in the following way. For each token of the expansion being analyzed, we examined the distance in words to the previous token of the expansion, measured between the first words of the two tokens. (Following the tokenization of items in the corpus, we treated punctuation symbols as words.) We excluded any token using exactly the same words as the previous token; in such cases, close repetition might arise simply because a certain phrase was the topic of discussion. We binned the intervals between tokens into categories at increments of powers of 10: 1–10, 11–100, and so on. (In fact, the minimum possible distance is two or three, depending on the size of the constituent; we incorporated this into our calculations but say no more about it here.) We define a repetition at a distance of <= 100 words a 'close repetition'; a repetition within 10 words is a 'very close repetition', and a repetition within 11 to 100 words is a 'somewhat close repetition'. Our focus is on very close repetitions: our intuition was that close repetitions sometimes happen at very short distances, and examples of rhetorical uses of repetition typically involve such distances as well. To a first approximation, an expansion's repetition tendency could be defined as the count of its very close repetitions as a proportion of its overall count. One issue here is that constructions might be repeated simply because they are favored by a particular author or happen to be appropriate for the subject matter. (For example, in a discussion of tax policy, one might repeatedly use phrases similar to *the rich*: *the poor*, *the wealthy, the disadvantaged*, etc.) The average length of sections (passages of text taken from a single source, such as a newspaper article or web page) within the COCA corpus is 2,436 words. Thus, repetitions within sections might give rise to repetitions on the order of 100 or 1,000 words. To reduce the effect of mere within-section repetition, we examined the ratio between very close repetitions (<= 10 words) and somewhat close repetitions (between 11 and 100 words).

Even without any preference for repetition, the measure just described might differ depending on the frequency of the expansion. If an expansion were extremely common, a random distribution might produce a high proportion of very close repetitions. Thus, for each expansion, we compared the measures just described to what would be expected if tokens of the expansion were randomly distributed. As noted by Myslín and Levy (2016), if tokens are randomly distributed, the intervals between them will follow a geometric distribution. The proportion of intervals falling within a certain range is the cumulative geometric distribution, which can be calculated analytically: if the count of an expansion is $c$ and its overall probability (its count divided by the number of words in the corpus) is $p$, the expected count of intervals $EC_{j,k}$ falling within the range $(j,k)$ (inclusive) is

$$EC_{j,k} = ((1 - (1-p)^{k}) - (1 - (1-p)^{(j-1)})) \cdot c$$

Our second aim was to determine the tendency for each NP expansion to be repeated in coordinate constructions. Following previous studies (Dubey et al., 2008; Temperley & Gildea, 2005), we considered just coordinate structures with two conjuncts, e.g. *the dog and the cat*. Coordinating conjunctions are labeled in the corpus as CC; the vast majority of these tokens are the word *and* (86.7%), with most of the rest being *or* (11.9%). In addition, more than 99% of the tokens of *and* and *or* are tagged as CC. Since it was more convenient at this stage of the process to work with words rather than preterminals, we took the words *and* and *or* as proxies for CC. Ideally, following previous studies (Dubey et al., 2008; Temperley & Gildea, 2015), we would examine all NP coordinates in which a certain expansion occurred in the first conjunct and the proportion of these in which the same expansion occurred in the second conjunct ("matching" coordinate structures). This was not possible, however; since the corpus does not show constituents, there is no way to identify and count NP coordinate structures in general. Instead, we simply examine the number of matching coordinate tokens as a proportion of the number of very close repetitions. Note that in matching coordinate constructions, the distance between the two NPs will always be either three (if the expansion is two words long, e.g. *the rich and the poor*) or four (if the expansion is three words long); thus these will always be counted as very close repetitions.

## 4. Results

The second column of Table 3 shows the number of valid intervals (intervals between consecutive tokens, excluding tokens that were identical to the previous token) for each of the fifteen expansion types. The five expansion types expected to be rare are indeed much lower in frequency than the types predicted to be common, though there is great variation both among the rare ones and among the common ones. Ideally one would define each frequency as a conditional probability given an NP parent; this is not possible to do precisely, since there is no good way of counting NPs in the corpus. In the WSJ corpus, there are about 380,000 NPs in 1.15 million words; this yields an estimate of 3.8 million NPs for the COCA corpus. With that estimate, the conditional probabilities of these expansions range from .00001 for <*a* $N_{prop}$ $N_{prop}$> to .08 for <D $N_{sing}$>. Since all of the expansions considered here have an NP parent, their raw frequencies in the COCA corpus are proportional to their conditional probabilities; thus raw frequencies are sufficient for present purposes.

**Table 3.** Repetition in rare and common NP expansions

| **Expansion** | **# valid intervals** | **Distance <= 10 (very close)** | **11 <= distance < 100 (somewhat close)** | **Very close (vc)/ somewhat close (sc) ratio** | **Observed/ expected vc/ sc ratio (*repetition tendency*)** | **Observed/ expected sc count** | **Proportion in coordinate constructions** |
|---|---|---|---|---|---|---|---|
| **Rare expansions** | | | | | | | |
| <$N_{sing}$ $N_{prop}$ $N_{prop}$> | 654 | 31 | 48 | 0.646 | 7.245 | 14.405 | 0.387 |
| <*a* $N_{prop}$ $N_{prop}$> | 49 | 5 | $1^{**}$ | 5.000 | 56.238 | 53.322 | 1.000 |
| <*the* A> | 1,489 | 199 | 78 | 2.551 | 25.350 | 4.533 | 0.246 |
| <D $N_{pl}$ $N_{sing}$> | 275 | 1 | 5 | 0.200 | 2.247 | 8.473 | 0.000 |
| <$N_{sing}$ A $N_{sing}$> | 63 | 1 | 3 | 0.333 | 3.749 | 96.776 | 1.000 |
| **Common expansions** | | | | | | | |
| <D $N_{sing}$> | 300,643 | 77,395 | 200,068 | 0.387 | 1.244 | 0.931 | 0.017 |
| <$N_{prop}$ $N_{prop}$> | 38,152 | 6,813 | 13,689 | 0.498 | 4.234 | 1.424 | 0.131 |
| <D A $N_{sing}$> | 96,659 | 8,981 | 50,120 | 0.179 | 1.353 | 1.028 | 0.027 |
| <A $N_{pl}$> | 55,313 | 6,041 | 22,966 | 0.263 | 2.084 | 1.224 | 0.081 |
| <D $N_{pl}$> | 45,367 | 3,354 | 15,324 | 0.219 | 1.807 | 1.163 | 0.046 |

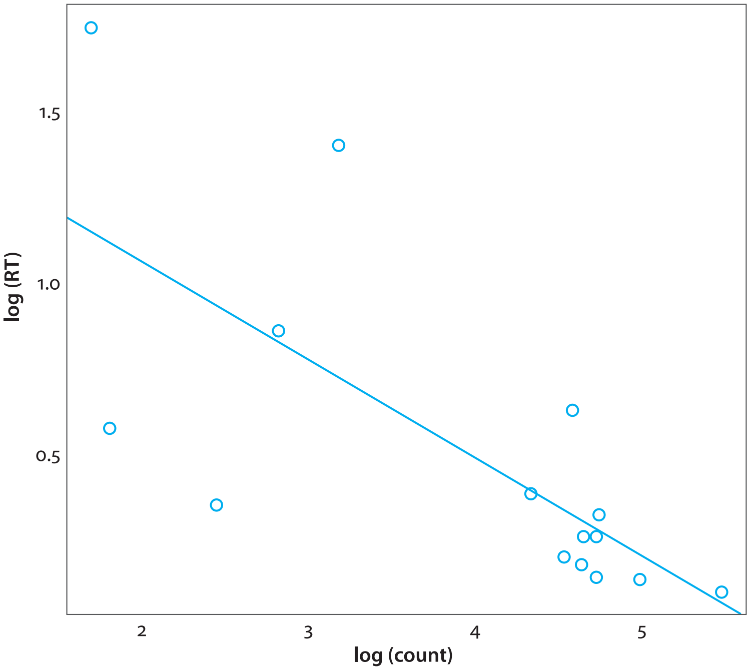
**Table 3.** *(continued)*

| Expansion | # valid intervals | Distance <= 10 (very close) | 11 <= distance < 100 (somewhat close) | Very close (vc)/ somewhat close (sc) ratio | Observed/ expected vc/ sc ratio (*repetition tendency*) | Observed/ expected sc count | Proportion in coordinate constructions |
|---|---|---|---|---|---|---|---|
| <A N$_{sing}$> | 52,282 | 4,473 | 19,924 | 0.225 | 1.801 | 1.174 | 0.106 |
| <N$_{sing}$ N$_{pl}$> | 21,342 | 1,423 | 5,403 | 0.263 | 2.405 | 1.669 | 0.145 |
| <D N$_{sing}$ N$_{sing}$> | 34,319 | 1,573 | 9,725 | 0.162 | 1.575 | 1.223 | 0.041 |
| <*the* A N$_{sing}$> | 43,546 | 2,410 | 15,142 | 0.159 | 1.491 | 1.230 | 0.031 |
| <*a* A N$_{sing}$> | 52,602 | 2,852 | 18,716 | 0.152 | 1.376 | 1.082 | 0.048 |

** The count for the range 11–100 was zero; in this case, we use the count for the range 101–1,000.

The third column in Table 3 shows the number of very close repetitions (<= 10 words) for each expansion type. These numbers (even considered as proportions) are not very meaningful, for two reasons mentioned earlier. First of all, some expansions may tend to be repeated within a section of the corpus (because they are favored by the author or appropriate for the topic), which would increase the number of very close repetitions. Secondly, common expansions may have a high proportion of very close repetitions even if tokens are randomly distributed. We can control, or at least reduce, both of these confounds by comparing the count of very close repetitions to somewhat close repetitions (between 11 and 100 words); the latter count is shown in the fourth column, and the ratio between very close and somewhat close repetition is shown in the fifth. Even this measure may be inflated for high-frequency expansions. To address this, we calculate the expected ratio of very close to somewhat close repetitions in a random distribution and examine the ratio between the observed "very-close-to-somewhat-close" ratio and the expected ratio; this is shown in the sixth column. We call this ratio the 'repetition tendency' of the expansion. A value of 1.0 would indicate that the expansion has no particular tendency toward very close repetition; a much higher value would indicate a strong tendency; a value less than 1.0 would indicate an avoidance of close repetition.

Figure 2 shows the log-transformed repetition tendency of each expansion plotted against its log-transformed frequency. Notably, all of the expansions have repetition tendencies greater than 1.0 (log > 0), even the common ones. Less frequent expansions generally show stronger repetition tendencies, as predicted by the inverse frequency effect; the correlation between log-transformed repetition tendency and log-transformed frequency is significant ($r = 0.29$, $p < .005$). This confirms the general pattern observed by Temperley and Gildea (2015) for NPs and by Reitter et al. (2006) for syntactic rules in general. However, the pattern is quite complex and uneven. All but one of the common expansions have repetition tendencies between 1.2 and 2.5; the exception is <$N_{prop}$ $N_{prop}$>, with a somewhat higher value of 4.234. Two of the rare expansions, <D $N_{pl}$ $N_{sing}$> and <$N_{sing}$ A $N_{sing}$>, also have low values. The value for <$N_{sing}$ $N_{prop}$ $N_{prop}$> is noticeably higher, and the values for <*a* $N_{prop}$ $N_{prop}$> and <*the* A> are much higher still. In fact, the value for <*a* $N_{prop}$ $N_{prop}$> is probably being underestimated. In this case, there were no tokens of somewhat close repetitions, so the "very-close-to-somewhat-close" ratio would be infinite. Instead, for the count of somewhat close repetitions, we used the range 101–1,000, which contains one token. In the discussion section, we discuss some possible reasons for these differences between rare expansions.



*Note.* For the fifteen expansions studied, the figure shows the (log) repetition tendency (the expected-to-observed ratio of the "very-close-to-somewhat-close" ratio) against the (log) count (base 10 is used in both cases).

**Figure 2.** Frequency and repetition tendency of NP expansions

The cutoffs defining very close and somewhat close repetition are somewhat arbitrary and give only an incomplete picture of the data. One might wonder whether the preference for close repetition tapers off gradually as distance increases or reflects a more sharply defined window. To investigate this, we examined the distribution of distances for one expansion, <*the* A>, in a more fine-grained manner. We chose to focus on this expansion because (i) it is one of two expansions (along with <*a* $N_{prop}$ $N_{prop}$>) that shows a very strong repetition tendency, and (ii) it has a much larger count than <*a* $N_{prop}$ $N_{prop}$> and thus seems to provide a better indicator of the "true" distribution that would be produced if unlimited data were available. Figure 3 shows the count of repetitions of <*the* A> at each distance from three to 20. (A distance of one is impossible, since the two tokens would overlap; a distance of two seems syntactically unlikely – the two tokens would have to be directly adjacent – and no cases of this were found in the corpus.) The counts decrease more or less monotonically as distance increases (disregarding some small fluctuations that seem likely to be due to chance). At all distances from three to six, counts are noticeably higher than at any larger distance; among distances of seven or greater, differences are small and inconsistent. Thus there seems to be a preference for repetition within a window of six words, at least for this expansion. One factor that may be at work here is a preference for repetition within a sentence rather than between sentences, which would naturally favor repetitions at shorter distances. This cannot easily be explored in the current corpus, since sentence breaks are not explicitly marked.

As noted earlier, it seemed likely that there would be some tendency toward syntactic repetition at distances outside the "very close" range, due perhaps to the subject matter or to the habits of particular authors. To explore this, we can examine the prevalence of "somewhat close" repetition (at distances between 11 and 100 words) in itself, again taking the ratio between the observed counts and those yielded by a random distribution; this is shown in the seventh column of Table 3. The values for <*a* $N_{prop}$ $N_{prop}$>, <D $N_{pl}$ $N_{sing}$>, and <$N_{sing}$ A $N_{sing}$> should be taken with caution, since the counts are so low (recall that the count for <*a* $N_{prop}$ $N_{prop}$> was adjusted from zero to one). For the remaining two rare expansions, <$N_{sing}$ $N_{prop}$ $N_{prop}$> and <*the* A>, the values are somewhat higher than the values for the common expansions, most of which are close to one. Even at moderate distances, then, rare expansions seem to show some tendency for repetition; common expansions show little or no such tendency. This suggests that rare expansions might be associated with certain specific topics or kinds of discourse; we return to this point in the discussion section.

The final column of Table 3 shows the proportion of very close repetitions that are in NP coordinate phrases: *X and Y*, where both *X* and *Y* are the same NP expansion. One expansion, <*a* $N_{prop}$ $N_{prop}$>, shows a strong bias toward coordinate repetition: all five tokens of very close repetition are in coordinate phrases.

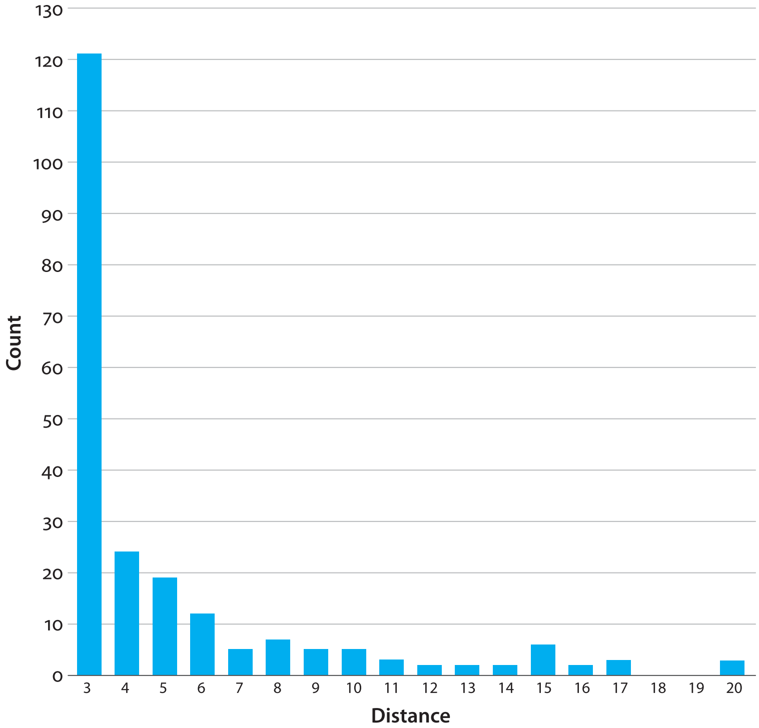


**Figure 3.** Counts of repetitions of <*the* A> at different distances (three through 20) in the COCA corpus

(For the expansions <D $N_{pl}$ $N_{sing}$> and <$N_{sing}$ A $N_{sing}$>, there was just one very close repetition; given so little data, it seems best to disregard these cases.) In all the remaining cases (both rare and common expansions), the proportion is below .4; in most cases it is much lower. This suggests that, on the whole, the repetition tendencies of these expansions are not primarily due to coordinate phrases. This surprised us, since Temperley and Gildea's (2015) study of repetition in NPs in general showed a much higher repetition tendency in coordinate phrases than in other situations (though they considered only cases where the two phrases were separated by one word). Particularly noteworthy is <*the* A>; like <*a* $N_{prop}$ $N_{prop}$>, this expansion has a very high repetition tendency, but in this case the proportion of tokens occurring in coordinate phrases is just .246. While this value is much higher than the corresponding values for any of the common expansions, it still suggests that most close repetitions of this construction do not occur in coordinate phrases. Out of the 121 repetitions of <*the* A> at a distance of three (see Figure 3), only 49 are in coordinate phrases. Inspection of the data suggested, however, that these values should be interpreted with caution. The reason is that the current measure only considers one specific kind of coordination. Examples (3a–e) below show several tokens of <*the* A> in the corpus that (arguably) involve kinds of coordination not captured by our measure. In (3a), three tokens of the expansion are conjoined in a coordinate phrase. Example (3b) has the feel of a coordinate phrase, but there is no coordinating conjunction (known as 'asyndetic coordination'). Similarly, (3c) has no conjunction, but the phrase *as well as* seems to function like one (one could easily substitute *and* for *as well as*). In (3d), there is a kind of coordinate parallelism, but it is not directly between the NPs; rather, it is between the clauses that contain them. On the other hand, many cases of very close repetition are not in any sense coordinate constructions – for example, where the expansion occurs in both the subject and object NPs of a clause, as in (3e).

(3) a. her distate for *the timid*, *the dull*, and *the ordinary*
b. the Mob generally only targets, keep in mind, *the successful*, *the prosperous*, *the influential.*
c. readers must thoughtfully attend to both sign systems – *the visual* as well as *the verbal* – ...
d. *the ugly* could become beautiful, and *the beautiful* could become plain
e. *the good* outweighs *the bad*

To explore this further, we categorized the very close repetitions of <*the* A> by hand, grouping them into five categories as shown in Table 4 (one might question whether or not "high-level coordination" should be considered a kind of coordination, but we do so here). We now see that a majority of very close repetitions – 134 out of 199 tokens – could be said to occur in coordinate constructions of some kind. However, there are still a substantial number that do not. Excluding the 134 coordinate tokens and re-running the test above, we found that the <*the* Adj> construction shows a repetition tendency of 8.030. Even without coordinate tokens, then, <*the* Adj> is used in very close repetition much more often than would occur by chance; thus, the phenomenon observed here cannot be attributed entirely to coordinate structures.

**Table 4.** Very close repetitions of the <*the* Adj> construction

| Syntactic environment | Count |
|---|---|
| Coordinate phrases with two conjuncts (includes 49 tokens found by automatic count as well as six tokens that were missed by that count) | 55 |
| Coordinate phrases with more than two conjuncts | 17 |
| Asyndetic coordinate phrases | 31 |
| High-level coordination – where the two tokens have the same syntactic role in larger conjoined phrases, e.g. both are subjects in conjoined clauses as in Example (3d) | 31 |
| Non-coordinate constructions | 65 |

## 5. Discussion

Earlier corpus studies have suggested that low-frequency syntactic constructions have an especially strong tendency to be repeated. In the current study, we examined the repetition tendencies of five rare NP expansions and ten common ones. Consistent with earlier studies (Reitter et al., 2006; Temperley & Gildea, 2015), we found that rare NP expansions show stronger repetition tendencies than common ones. However, there is great variation in repetition tendency among the rare expansions; two of them, <D $N_{pl}$ $N_{sing}$> and <$N_{sing}$ A $N_{sing}$>, show repetition tendencies similar to those of common expansions, while one, <$N_{sing}$ $N_{prop}$ $N_{prop}$>, has a somewhat stronger tendency and two, <*the* A> and <*a* $N_{prop}$ $N_{prop}$>, have much stronger tendencies. In this section we consider possible explanations for this variation and discuss some possible limitations of our study.

### 5.1 Explaining the results

How can the variation in repetition tendency among rare expansions be explained? We believe the answer may lie partly in the type of discourse in which these expansions are used, and in particular, the distinction between persuasive and informative discourse (Brewer, 1980; Sanders, 1997). Examples (4a–d) below show two examples of each of the two highly repetitive expansions, <*the* A> and <*a* $N_{prop}$ $N_{prop}$>. These examples are rather typical of their general use; all four feature very close repetitions.

(4) a. Today the church ignores *the afflicted*, comforts *the comfortable*, …
b. The ancient Greeks favored *the beautiful* over *the expressive*.
c. And that, for me, is a fundamental difference between *a Ronald Reagan* and *a Barack Obama*.
d. When I graduated from college … I didn't have *a Hank Aaron* or *a Willie Mays* to talk about.

The first example is emotionally charged: clearly the author has a negative view of "the church" and wishes to win readers over to this position. The third example is similar, implying a strong preference between the two presidents (though it is not obvious which one is preferred). The second example, while less overtly emotional, is of a subjective nature – an opinion rather than an objective, testable claim. The fourth case seems designed to elicit sympathy for the author's lack of African-American role models in baseball. All four of the examples could be described as persuasive discourse, or in other words, as rhetorical. They advocate a particular view of a topic, and aim to convince readers of this view, rather than simply describing a situation or presenting facts in an objective way – as one would find, for example, in a "hard news" article, a scientific paper, or an instruction manual. Returning to Table 3 ("observed/expected sc count"), the fact that <*the* A> has a fairly high tendency toward "somewhat close" repetition (much higher than that of any common construction) suggests that it is associated with certain kinds of discourse. (For <*a* $\mathrm{N}_{\mathrm{prop}}$ $\mathrm{N}_{\mathrm{prop}}$>, the count of somewhat close repetition was actually zero; given the very low overall frequency of this expansion, its tendency toward somewhat close repetition cannot be reliably estimated.) By contrast, the two rare constructions with low repetition tendencies examined here, <D $\mathrm{N}_{\mathrm{pl}}$ $\mathrm{N}_{\mathrm{sing}}$> and <$\mathrm{N}_{\mathrm{sing}}$ A $\mathrm{N}_{\mathrm{sing}}$>, seem much less associated with persuasive discourse. No doubt they *could* be used in such discourse, but many of their uses appear to be purely informative, as in Examples (5a–b):

(5) a. The company blamed a number of factors for *the earnings decline* ...
b. This provided the number of *core academic teachers* per student ...

In short, we suggest that constructions with high repetition tendencies tend to be those that are associated with persuasive discourse. This argument accords with the well-established idea that repetition is often used for rhetorical purposes (Corbett, 1971; Vickers, 1994). It also raises the further question of why certain expansions might come to be associated with persuasive discourse; we have no answer to this at present.

An interesting case is <$\mathrm{N}_{\mathrm{sing}}$ $\mathrm{N}_{\mathrm{prop}}$ $\mathrm{N}_{\mathrm{prop}}$>, whose repetition tendency is intermediate between the low and high values already discussed. In general, this construction does not appear to be strongly associated with persuasive discourse. Unlike a phrase like *the rich*, phrases like *architect Julia Morgan*, *singer Nancy Sinatra*, and *outfielder Brian Hunter* do not, in themselves, suggest rhetorical intent. We do note, however, that this construction tends to be used in less formal kinds of writing, and about topics that could be considered entertainment: in particular, writing about the arts and sports. For example, one would not expect a phrase like *singer Nancy Sinatra* to appear in an academic article about popular music or a "hard news" article about an incident at a concert. Like <*the* A>, this construction also shows a strong tendency toward "somewhat close" repetition (see Table 3), suggesting that it may be associated with certain kinds of discourse. Entertainment writing tends to be somewhat persuasive in character; an article about a baseball game might have a mixture of factual statements and strongly opinionated ones. If (for whatever reason) the <$\mathrm{N}_{\mathrm{sing}}$ $\mathrm{N}_{\mathrm{prop}}$ $\mathrm{N}_{\mathrm{prop}}$> construction is associated with entertainment writing, the opinionated style of that type of discourse might encourage a high repetition tendency.

The previous paragraph makes an important general point: the distinction between persuasive and informative discourse is a continuum. Scientific articles may sometimes feature discourse that is persuasive to some degree (one might cite the current discussion as an example!). Undoubtedly, the degree to which a construction is associated with one type of discourse or another is a continuum as well. For this reason, we put forth the performative/informative distinction as a tentative explanation for our findings, deserving further exploration. There is no obvious way of pursuing this issue further using the current corpus. While the corpus is divided into eight sections – academic, blog, fiction, magazine, news, spoken, TV/movies, and web – inspection suggests that each of them contains a mixture of informative and persuasive discourse. The sections do differ in other ways – in level of formality, and in the balance between spoken and written discourse; the effects of these factors on syntactic repetition would be interesting to consider, but we leave that as a project for the future.

One might ask whether the discourse-based account of syntactic repetition explains the inverse frequency effect. Do the constructions associated with persuasive discourse tend to be rare, and would we expect this to occur? One might say, if certain constructions are confined to a certain type of discourse, that alone could make them rare. On the other hand, as noted earlier, scholars of rhetoric have argued that the use of rare or marginal syntactic constructions is an effective rhetorical device in itself. In any case, if (for whatever reason) both syntactic repetition and rare constructions are associated with persuasive discourse, then an association between syntactic repetition and rare constructions could arise through this indirect route, giving rise to the inverse frequency effect. However, we suspect that information flow also plays a role in the inverse frequency effect, independent of discourse type. Information flow makes further predictions about syntactic repetition that do not follow from a discourse-based account: that lexical probabilities in coordinate phrases will be lower in matching second conjuncts than first conjuncts, and lower in matching second conjuncts than in nonmatching ones (Temperley & Gildea, 2015). It seems likely, then, that both information-based and discourse-based factors affect syntactic repetition. It may be that information flow encourages repetition of rare constructions in general, but discourse-based pressures favor it in some rare constructions more than others.

One further explanation for the inverse frequency effect deserves consideration. Jaeger and Snider (2013) examine the dative alternation in a corpus of spontaneous dialogue and find that a particular form of the alternation that is unexpected in context (because it is disfavored for a particular verb) is more likely to be repeated. Jaeger and Snider suggest that this is because an unexpected use of a construction by one participant in a dialogue causes a "prediction error" on the part of the other participant, making them favor that construction in the near future. Jaeger and Snider's account may be compatible with the account of syntactic repetition advanced here. By either a "prediction error" account or an "information flow" account, repetition increases the probability of an expansion, and thus facilitates its processing in a way that is amplified for rare constructions. The strong repetition tendencies of (some) rare constructions, even within the output of a single writer or speaker, may be a response to this facilitative effect – an aspect of "audience design" (Ferreira, 2019). However, Jaeger and Snider use conversational data; such data is very different in character from the data used here, most of which reflects output of a single writer or speaker. The similarities and differences in the use of syntactic repetition between conversational and non-conversational discourse deserve further exploration.

## 5.2 Limitations

The current study has several methodological limitations. The process of identifying tokens of expansions may be subject to both misses and false positives. For the rare expansions, there should be few false positives, since tokens were hand-checked (though there might be a small number due to human error in the checking process). Misses could occur if valid tokens were missed by the automatic search process, due to incorrect preterminal tags or to heuristics used to limit the number of tokens. One test of this was reported earlier, validating our heuristic assumption that <*the* A> will very rarely be followed by a noun. As another small test, we searched the corpus manually for tokens of <$N_{sing}$ $N_{prop}$ $N_{prop}$>, using simple string searching (three words, the first beginning with a lower-case letter and second and third with capital letters). Of the first ten valid tokens we found, eight were also found by the automatic search process; the two misses were both due to incorrect labeling of preterminals. As noted earlier, due to the statistical nature of tagging systems, incorrect preterminal labels may well be more common with rare constructions.

Another limitation of our study concerns the way expansions are defined. We define each expansion as a sequence of two or three preterminals (and/or words). Consider the restrictive appositive construction, <$N_{sing}$ $N_{prop}$ $N_{prop}$>. Our search process allows phrases like *trumpeter Wynton Marsalis*, but it disallows *opposition leader Joseph Rendjambe*, since in that case the initial common noun has a noun modifier, and *historian Samuel Eliot Morison*, because the proper name contains three words. It is possible that, with regard to the preference for repetition, all of these are regarded as instances of the same construction (another possible variant includes a determiner, e.g. *the astronomer Edmund Halley*, but this seems to us like a qualitatively different construction.) To examine this, we reran the test on <$N_{sing}$ $N_{prop}$ $N_{prop}$>, including phrases with any number of singular or proper nouns (but excluding those with determiners). We found that the repetition tendency was somewhat lower than in the original test: 3.6 versus 7.2. This might be taken as evidence that the cognitive representation of the restrictive appositive construction is sensitive to length, and that repetitions involving tokens of similar length are especially favored. This does not mean, however, that our original definition of the expansion is optimal; there are other possibilities, such as allowing different lengths in the proper noun portion but not the common noun portion. Similar issues arise with the other expansions considered here. We doubt that this problem could explain away the very large differences we observe between expansions in repetition tendency, though it is difficult to be certain.

## 6. Conclusions and future directions

In the current study, we have offered an exploratory investigation of the inverse frequency effect – the tendency for low-frequency syntactic constructions to be used repetitively. We examined five low-frequency NP expansions in written English and their tendency to be repeated at short distances. Three of the expansions showed strong repetition tendencies (compared to common NP expansions); the other two did not. We suggest that this difference among rare syntactic constructions may be due to their associations with different kinds of discourse: the constructions showing strong repetition tendencies are associated with persuasive rather than informative discourse. Seen in this way, the repetitive use of certain constructions could be seen as an example (along with others previously observed) of the use of repetition for rhetorical purposes. Given the small number of constructions considered here, our explanation is conjectural, requiring further confirmation. We also do not claim that discourse type is the only factor influencing syntactic repetition; we suspect that uniform information density also plays a role (repeating low-frequency constructions smooths out the flow of information). But discourse type is an additional factor that should be taken into account in future studies of syntactic repetition.

An interesting situation not considered so far is illustrated by Examples (6a–b):

(6) a. Then we have *comedian and actor Rick Younger* ...
b. ... where *the rich and famous* build their mansions ...

Each of the examples resembles one of the rare NP expansions discussed earlier (<$\mathrm{N}_{\mathrm{sing}}$ $\mathrm{N}_{\mathrm{prop}}$ $\mathrm{N}_{\mathrm{prop}}$> in the first case, <*the* A> in the second); one might say that there is repetition *within* the expansions. It is not obvious, however, that these cases represent repetition of a syntactic pattern – at least, a *rare* syntactic pattern. In terms of constituent structure (assuming the Penn Treebank annotation system), all that is being repeated is a single preterminal type (singular noun or adjective), and of course these types are not rare at all. A richer constituent representation could perhaps indicate a rare, repeated structure – for example, if each adjective in (6b) was followed by an empty N. This could also be accomplished by viewing the phrases as dependency structures, such that the syntactic use of a word is represented by its dependency relations with other words. Example (6a) features two singular common nouns, each of which is a dependent of the following proper noun phrase; (6b) features two adjectives that are presumably connected to the previous determiner and not to a following noun (whether the adjectives are heads or dependents of the determiner is not important for present purposes). These are unusual uses of common nouns and adjectives, respectively. Thus, the inverse frequency effect might apply to such cases: we would expect rare syntactic word usages to have strong repetition tendencies. For example, we might expect a stronger repetition tendency for rare uses of singular nouns (a high frequency of (7a) below in proportion to Example (7b)) than for more common uses such as <D $N_{sing}$> ((7c) in proportion to Example (7d)). This prediction has not been tested.

(7) a. $N_{sing}$ *and* $N_{sing}$ $N_{prop}$ $N_{prop}$ (e.g. *comedian and actor Rick Younger*)
b. $N_{sing}$ $N_{prop}$ $N_{prop}$ (*comedian Rick Younger)*
c. D $N_{sing}$ *and* $N_{sing}$ (*the dog and cat)*
d. D $N_{sing}$ (*the dog*)

A connection arises here with a study by Heycock and Zamparelli (2003) on the use of 'bare' count nouns (i.e. those without a determiner). Bare count nouns are generally ungrammatical in English, but they seem much more felicitous in coordinate structures. Examples (8a–b) and (9a–b) are from Heycock and Zamparelli; (8) is the context for (8a) and (8b); (9b) is not explicit in their paper, but implied.

(8) A black cat and a brown dog were fighting in the street.
a. Cat and dog were equally filthy.
b. *Cat was filthy.

(9) a. At the company meeting, president and vice president gave an optimistic speech.
b. *At the company meeting, president gave an optimistic speech.

By the reasoning presented in the previous paragraph, the bare count noun could be considered a rare syntactic pattern that has a high repetition tendency. One might say the situation is somewhat different from those considered earlier, since bare count nouns used singly are not just rare but ungrammatical. Even so, this could be regarded as an extreme case of the inverse frequency effect, and Heycock and Zamparelli's frequent use of intermediate grammaticality symbols such as *? and ?? suggests that they find some gradience in this area. A corpus study of this phenomenon would be worthwhile, though difficult, since neither the Penn Treebank nor the COCA corpus distinguishes count nouns from mass nouns. (This, too, is a gradient distinction; *cheese* is normally a mass noun, but one can speak of *a delicious cheese.*) Searching the COCA corpus for patterns of the form $\text{N}_{\text{sing}}$ *and* $\text{N}_{\text{sing}}$ (not preceded by a determiner), we do find occasional cases of coordinated bare count nouns, such as those below. Confirming Heycock and Zamparelli's intuition, Examples (10a–d) seem much less felicitous if only one noun is used (e.g. **she was neatly attired in skirt*).

(10) a. ... she was neatly attired in skirt and blouse, ...
b. The deep blue of the skies was reflected in loch and river, ...
c. Maggot hunted with spear and arrow, ...
d. The canvas ... is noteworthy for its varied treatment of face and gesture ...
e. This allows both student and teacher to chart progress through an assignment ...

Can close repetition improve bare count nouns in other situations besides coordinate structures? Jackendoff (2008) discusses expressions of the form NPN, such as those in Examples (11a–e). Such expressions feature two bare count nouns in close proximity. While many could be considered idiomatic, the construction can also be used productively, as in (11d) and (11e). Unlike other constructions considered here, however, the NPN construction often features repetition of the noun, as in (11c–e).

(11) a. cheek by jowl
b. hand over fist
c. day to day (or, from day to day)
d. we read book after book
e. We examined the moon crater by crater

Whether coordinated bare count nouns and the NPN construction can be considered instances of the inverse frequency effect may be debatable, but there is at least a suggestive connection between them and the rare constructions explored in earlier sections. In all of these cases, a rare construction or syntactic usage seems to be licensed or improved by repetition – or at least, to have a tendency toward repetition, suggesting that repetition has a desirable effect. This leads to the further prediction that coordinated bare count noun constructions and the NPN construction might be especially associated with persuasive rather than informative discourse. Examples presented in this section suggest to us that they may be, but this requires further study.

## References

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31–56.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*(3), 355–387.

Brewer, W. F. (1980). Literary theory, rhetoric, and stylistics: Implications for psychology. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical Issues in Reading Comprehension* (pp. 221–239). Erlbaum.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*(2), 234–272.

Charniak, E., Blaheta, D., Ge, N., Hall, K., Hale, J., & Johnson, M. (2000). Bllip 1987–89 WSJ corpus release 1. *Linguistic Data Consortium*, *36*.

Cleland, A. A., & Pickering, M. J. (2003). The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language*, *49*(2), 214–230.

Corbett, E. (1971). *Classical Rhetoric for the Modern Student (2nd Edition)*. Oxford University Press.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159–190.

Dubey, A., Keller, F., & Sturt, P. (2008). A probabilistic corpus-based model of syntactic parallelism. *Cognition*, *109*(3), 326–344.

Fenk, A., & Fenk, G. (1980). Konstanz im Kurzzeitgedächtnis – Konstanz im sprachlichen Informationsfluß? *Zeitschrift für Experimentelle und Angewandte Psychologie*, *27*(3), 400–414.

Ferreira, V. S. (2003). The persistence of optional complementizer production: Why saying "that" is not saying "that" at all. *Journal of Memory and Language*, *48*(2), 379–398.

Ferreira, V. S. (2019). A mechanistic framework for explaining audience design in language production. *Annual Review of Psychology*, *70*(1), 29–51.

Frank, A., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Meeting of the Cognitive Science Society* (pp. 939–944). https://escholarship.org/uc/item/7d08h6j4

Frazier, L., Munn, A., & Clifton, C. (2000). Processing coordinate structures. *Journal of Psycholinguistic Research*, *29*(4), 343–370.

Garside, R., & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech, & A. McEnery, A. (Eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora* (pp. 102–121). Longman.

Gries, S. T. (2005). Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research*, *34*(4), 365–399.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In In R. Levy & R. Reitter (Eds.), *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics* (Vol 2., pp. 159–166). Association for Computational Linguistics.

Hartsuiker, R.J., Kolk, H.H., & Huiskamp, P. (1999). Priming word order in sentence production. *The Quarterly Journal of Experimental Psychology Section A*, *52*(1), 129–147.

Heycock, C., & Zamparelli, R. (2003). Coordinated bare definites. *Linguistic Inquiry*, *34*(3), 443–469.

Huddleston, R., & Pullum, G.K. (2002). *The Cambridge Grammar of the English Language*. Cambridge University Press.

Jackendoff, R. (2008). Construction after construction and its theoretical challenges. *Language*, *84*(1), 8–28.

Jaeger, T.F., & Snider, N.E. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the primes prediction error given both prior and recent experience. *Cognition*, *127*(1), 57–83.

Konopka, A.E., & Bock, K. (2009). Lexical or syntactic control of sentence formulation? Structural generalizations from idiom production. *Cognitive Psychology*, *58*(1), 68–101.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Levy, R., & Jaeger, T.F. (2007). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt, & T. Hofmann (Eds.), *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference* (pp. 849–856). MIT Press.

Marcus, M.P., Kim, G., Marcinkiewicz, M.A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., & Schasberger, B. (1994). The Penn Treebank: Annotating predicate argument structure. In *ARPA Proceedings of the Workshop on Human Language Technology* (pp. 114–119). Morgan Kaufmann.

McQuarrie, E.F., & Mick, D.G. (1996). Figures of rhetoric in advertising language. *Journal of Consumer Research*, *22*(4), 424–438.

Menninghaus, W., Wagner, V., Wassiliwizky, E., Jacobsen, T., & Knoop, C.A. (2017). The emotional and aesthetic powers of parallelistic diction. *Poetics*, *63*, 47–59.

Myslín, M., & Levy, R. (2016). Comprehension priming as rational expectation for repetition: Evidence from syntactic processing. *Cognition*, *147*, 29–56.

Pickering, M.J., & Branigan, H.P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, *39*(4), 633–651.

Pickering, M.J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190.

Quirk, R., Greenbaum, S., Leech, G., & Svartik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman.

Reitter, D., Keller, F., & Moore, J.D. (2006). Computational Modelling of Structural Priming in Dialogue. In R.C. Moore, J. Bilmes, J. Chu-Carroll, & M. Sanderson (Eds.), *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 121–124). Association for Computational Linguistics. https://aclanthology.org/N06-2031.

Reitter, D., Keller, F., & Moore, J.D. (2011). A computational cognitive model of syntactic priming. *Cognitive Science*, *35*(4), 587–637.

Sanders, T. (1997). Semantic and pragmatic sources of coherence: On the categorization of coherence relations in context, *Discourse Processes*, *24*(1), 119–147.

Scheepers, C. (2003). Syntactic priming of relative clause attachments: Persistence of structural configuration in sentence production. *Cognition*, *89*(3), 179–205.

Szmrecsanyi, B. (2005). Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory*, *1*(1), 113–150.

Temperley, D. (2022). Music and language. *Annual Review of Linguistics*, *8*, 153–170.

Temperley, D., & Gildea, D. (2015). Information density and syntactic repetition. *Cognitive Science*, *39*(8), 1802–1823.

Trueswell, J.C., & Tanenhaus, M.K. (1994). Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. In C. Clifton, L. Frazier, & K. Rayner (Eds.), *Perspectives on Sentence Processing* (pp. 155–179). Lawrence Erlbaum Assoc.

Vickers, B. (1994). Repetition and emphasis in rhetoric: Theory and practice. In A. Fischer (Ed.), *Repetition (Swiss Papers in English Language and Literature, 7)* (pp. 85–114). Gunter Narr Verlag.

## Appendix A. Identifying NP expansions

### I. *The corpus and general procedure*

Here we explain our procedure for identifying tokens of rare and common NP expansions.

A few comments are needed about our corpus, an 11-million-word sample of the COCA corpus ("words" here includes punctuation symbols.) The sample was downloaded from https://www.corpusdata.org/formats.asp on 8/11/21. The corpus contains a series of lines, each with the format *textblock# word lemma preterminal*; only the words and preterminals (part-of-speech tags) were used here. Once in every 200 words, a sequence of ten words (and tags) is blocked out (replaced with "@") for reasons of "fair use"; however, it is possible to identify the missing words by finding the passage in the online, full COCA corpus (https://www.english-corpora.org/coca/), and we sometimes did so during the hand-filtering process. Some words have multiple tags, e.g. **nn1_jj**;[1] this seems to indicate uncertainty in the tagging process (the word could be a singular noun or an adjective). Generally we took the first tag in such cases. The only exception was <A $N_{\text{prop}}$ $N_{\text{prop}}$>; because this expansion was so rare, we allowed the proper-noun tag **np** anywhere in the preterminal tags, so as to gather more tokens.

As explained in the main text, all of the desired expansion types are "flat", i.e. not containing any phrasal constituents, and two or three words in length. In identifying NPs, we tried to use the same criteria used in the Penn Treebank. Generally, each flat NP may contain only one head noun; a phrase such as *royalty and rock stars* is treated as two conjoined NPs. However, two or more conjoined nouns with no modifiers (e.g. *champagne and dessert*) are treated as a single flat NP, as are multiple nouns modified by the same modifier (*future trade and investment*). Noun modifiers of a single head remain within the flat NP, even if they are conjoined or have their own modifiers, e.g. *the eye care and skin care concern*. Possessive lexical NPs (not pronouns) always generate their own NPs, e.g. (NP (NP *John's*) *dog*). These criteria affect our tokenization process in ways that may not always be obvious. For example, in the phrase *lung cancer deaths*, *lung cancer* is not treated as a token of <$N_{\text{sing}}$ $N_{\text{sing}}$> because (by Penn Treebank conventions) it is part of a larger flat NP. Possessive NPs were allowed as valid tokens (e.g. in (NP (NP *missionary David Brainerd's*) *diary*), *missionary David Brainerd* is a token

[1] Preterminal tags will be shown here in boldface, words in italics, and expansions surrounded by angle brackets.

of <$N_{sing}$ $N_{prop}$ $N_{prop}$>), though strictly speaking they should not be because the possessive marker would be included in the NP.

For each expansion, we began by identifying possible tokens using an automatic process. In the case of rare expansions, we then hand-filtered the output of the search to exclude false positives; for common expansions, no hand-filtering was done. Our automatic search processes consider both preterminals and words (with words, both initial-capitalized and uncapitalized forms are considered). They search for sequences of elements corresponding to the expansion of interest, but also consider the context, using heuristics to filter out false positives. A general concept that we found useful was 'possible NP beginners', which are preterminal types that often start NPs. These include the following (letters in parentheses indicate the first letter of the preterminal): articles and possessive pronouns (**a**); quantifiers, demonstratives, and wh-words (**d**); adjectives (**j**); numbers (**m**); and nouns (**n**). In searching for certain NP expansions (especially those that did not start with determiners), we excluded cases where the previous word was a possible NP beginner. No doubt this heuristic excluded some good tokens, but probably (we believe) very few. (As one test, for the <$N_{sing}$ A $N_{sing}$> expansion, we tried insisting that the previous word was one of these types. In the first 3 million words of the corpus, we found only two good tokens that were produced by this search and thus were missed by excluding such tokens.) With most searches, we filtered out cases where the word following the target sequence was a noun, since the target sequence was most often a noun-modifier phrase.

In Sections II and III below, we provide further detail about the search process for the expansions. For each expansion, we indicate the number of valid tokens found. As explained in the main text, our analysis of the data examined the intervals (distances in words) between consecutive tokens of an expansion. However, we disregarded any token that was identical in words to the previous token. Thus, the number of valid *intervals* may be less than the number of valid tokens minus 1. (The number of valid intervals for each expansion can be seen in Tables 2 and 3 of the paper.) In some cases, this exclusion significantly affects the results. For example, one long article in the corpus is on the topic of "the sublime," and contains many consecutive tokens of <*the* A> using this phrase.

## II. *Common expansions*

For common expansions, we simply applied an automatic search process with no hand-filtering. In what follows, we indicate how each common construction was defined. Our labels for preterminals in the paper generally map on to labels in the corpus as follows: D = **at**, $N_{sing}$ = **nn1**, $N_{pl}$ = **nn2**, $N_{prop}$ = **np**, A = **jj**. Each preterminal label below refers to the initial portion of a label; so, **n** would match **nn1**, **nn1_jj**, etc. (one exception is that **jj** excludes **jjr** – comparative adjectives – and **jjt** – superlative adjectives). "[**adjmn**]" means any label beginning with any of those five letters. [*a*|*an*] means *a* or *an*. The number of valid tokens of each expansion is shown in parentheses.

1. <D $N_{sing}$> (311,140): **at nn1**, not followed by **n**
2. <$N_{prop}$ $N_{prop}$> (42,980): **nnp nnp**, not preceded by [**adjmn**], not followed by **n**
3. <D A $N_{sing}$> (98,631): **at jj nn1**, not followed by **n**
4. <A $N_{pl}$> (56,759): **jj nn2**, not preceded by [**adjmn**] or followed by **n**
5. <D $N_{pl}$> (49,190): **at nn2**, not followed by **n**
6. <A $N_{sing}$> (54,791): **jj nn1**, not preceded by [**adjmn**] or by **jj** [**c**,], not followed by **n**
7. <$N_{sing}$ $N_{pl}$> (22,710): **nn1 nn2**, not preceded by [**adjmn**] or by **n cc**, not followed by **n**

8. <D $N_{sing}$ $N_{sing}$> (36,186): **at nn1 nn1**, not followed by **n**
9. <*the* A $N_{sing}$> (45,222): *the* **jj nn1**, not followed by **n**
10. <[*a*|*an*] A $N_{sing}$> (53,619): *a/an* **jj nn1**, not followed by **n**

## III. *Rare expansions*

For the five rare expansions, we describe both the automatic process and the hand-filtering process. For the hand-filtering process, we examined a 20-word context around each possible token. In the released data (described in Appendix B), valid tokens of an expansion are marked *g* (good); invalid tokens are marked *x* or sometimes with other symbols.

1. <$N_{sing}$ $N_{prop}$ $N_{prop}$>

The automatic process looks for **nn**(uncapitalized) **np np**. This yielded 3,898 tokens. By excluding cases where the first word is capitalized, we avoid cases like *Senator Barbara Boxer* where the common noun is part of the title (though this causes sentence-initial tokens to be missed).

430 cases (marked *b*) were not restrictive appositives at all; in most of these, the singular noun was in a different phrase from the proper noun, e.g. *the impulse James Clifford calls Ethnographic Surrealism*. The remaining false positives were restrictive appositive phrases of some kind, but not matching our definition. In 577 cases (marked *d*) the restrictive appositive phrase had a determiner (*the astronomer Edmund Halley*). And in 2,228 cases (marked *x*), the phrase was singular and determinerless, but the wrong length, usually because there were either two or more common noun modifiers (e.g. *crowd favorite Clarence Clemons*), or three or more words in the proper name (*historian Samuel Eliot Morison*). This left 662 valid tokens. Nearly all cases seemed clear cut.

2. <*a* $N_{prop}$ $N_{prop}$>

Our automatic process found all two-word sequences of [*a*|*an*] **np** not followed by **nn** (or [*a*|*an*] **np np** not followed by **nn**). This is extremely permissive, since the construction actually requires two proper nouns. But similar constructions with a single proper noun do occur, e.g. *set in a Boston ruled by a mayor reminiscent of James Michael Curley*. While we exclude such cases, we thought that including them in the automatically generated output might be of value for future studies.

As we conceive of it, <*a* $N_{prop}$ $N_{prop}$> involves a proper name, referring to a single person (usually) or thing that is being recast as a category of similar people or things. *If there were a Thurgood Marshall on the bench* means *If there were someone similar to Thurgood Marshall on the bench*. Thus, in the hand-filtering stage, we excluded things like *a Saint Bernard*; *Saint Bernard* is already a category, not a single individual. The phrase *a Fra Angelico* was not included because (in context) it refers to a category of paintings by Fra Angelico rather than a category of people. One tricky case is a usage found in detective TV shows: *we got no record of a Darryl Kennedy*. This seems slightly different from the desired construction, since in this case, Darryl Kennedy represents a literal category of people (people named Darryl Kennedy), rather than people similar to a known person. Still, we included such cases; none were involved in very close repetitions.

The automatic search process found 1,454 tokens; 54 of these were judged to be valid tokens.

## 3. <*the* A>

Most often, <*the* A> has generic plural reference (*The new tax proposal will mostly benefit the rich*) and takes plural agreement (*The rich love the new proposal*). In such cases, it could be seen to substitute for a full noun phrase like *rich people* or *the rich people* (some might consider *the rich* to have indefinite plural reference, denoting not all rich people but an undefined subset of rich people: in that case *some rich people* would be a better substitute.) In other cases, <*the* A> seems to substitute for a mass noun phrase, like *The ancient Greeks favored the beautiful over the expressive*. Such uses seem to require singular agreement, e.g. *In Ancient Greece, the beautiful was/*were favored over the expressive)*, though it is not clear that the best substitute noun would be singular, or indeed, that any noun provides an adequate substitute (beautiful things? beautiful content? the beautiful realm?). Singular countable reference is rare for <*the* A>, and usually seems odd or ungrammatical (*There was a big pumpkin and a small one;* ?*we chose the big*). It does occasionally occur, in phrases like *So I pointed out the obvious* or *Then the inevitable happened*.

The automatic search process found cases of *the* **jj**. We excluded comparative (**jjr**) and superlative (**jjt**) adjectives, since these usually represent different constructions; comparative adjectives most often represent the correlative comparative (e.g. *the bigger the better*) and superlatives are commonly used without nouns, often in ways that seem more like predicative adjective or adverbial phrases (e.g. *he is the tallest and runs the fastest*). We also imposed several other restrictions. The vast majority of cases of *the* **jj** are false positives, in which the pattern actually forms the first part of a full noun phrase. To avoid these, we excluded cases where the pattern is followed directly by a noun (*the big dog*), an adjective (*the big black dog*), a number (*the warm 2012 winter*, *the late 1950s*), a comma or coordinate conjunction followed by an adjective (*the big, black dog*; *the big and black dog*), the words *one/ones*, or @ (the start of a blocked-out sequence of words).

Adjectives in <*the* A> cannot be pluralized: one cannot say *this tax policy favors the riches* or *the Greeks favored the beautifuls over the expressives*. Phrases with adjectives that seemed to allow pluralization were deemed not valid: four such words (*other*, *military*, *opposite*, and *original*) were excluded at the automatic stage (we would say *the militaries of Russia and the U.S. are very different*, not *the military of…*). The automatic search process also excluded several other common phrases. *The like* was excluded since *like* cannot usually be used as an attributive adjective at all (**the like dog*). *The potential* (most often followed by *to*) does not seem to substitute for any full noun phrase. *The past*, *the outside*, and *the left* were excluded for reasons of consistency: in the related phrases *the future*, *the inside*, and *the right*, the second word is labeled as a noun, so it seemed wrong to include the former phrases without including the latter.

The automatic procedure described above yielded 4,244 tokens. The vast majority of remaining false positives were clear-cut cases. In many cases, the pattern was part of a full noun phrase that was missed by the exclusions described above. For example, in the phrase *the absurd, historically great season*, *absurd* is clearly an attributive adjective modifying *season*; this was not excluded by the initial filtering since the following adjective *great* is preceded by an adverb. In some cases, the adjective was simply mislabeled: in the phrase *the fit and feel are awesome*, *fit* is labeled as an adjective but is obviously a noun (though the word can also act as an adjective).

As mentioned earlier, true tokens of <*the* A> seem to substitute for full noun phrases. Some phrases were excluded for this reason at the hand-filtering stage, including idioms such as *on* *the cheap* and *at the ready*; in such cases, no full noun phrase seems remotely appropriate as a replacement. This criterion was also useful with regard to passive and progressive participles, which are sometimes (though rather inconsistently) tagged as **jj**. Phrases with passive participles usually seem to qualify as tokens of <*the* A>, e.g. *the privileged* and *the disadvantaged*; in such cases a full noun phrase is clearly implied (privileged people). Progressive participles vary in this regard. In *comfort the dying*, there is an implied full noun phrase, but in *stop the bleeding*, there is not; *the bleeding* describes a process.

A small number of cases were difficult to classify. Some were idioms like *in the dark*, where it is unclear whether *the dark* substitutes for a noun phrase (we feel it does not). Others were cases where a "gapped" analysis was possible: that is, where the use of the pattern seemed to be licensed by the nearby presence of the implied noun. Consider these cases:

a. both the inner and the outer worlds
b. from the supine to the upright position
c. sorting the good policemen from the bad

In (a), *inner* is clearly an attributive adjective modifying *worlds* (though the repetition of the determiner is unusual). In (b), *supine* clearly modifies *position*, but treating it as attributive seems dubious as *position* is in a separate prepositional phrase. Still, the use of *supine* seems to be improved by the fact that *position* appears shortly afterwards. Generally, the use of <*the* A> does not require a nearby occurrence of the implied noun (though there might happen to be one). In (2c), the implied noun *policemen* appears shortly before *the bad*, but does not seem necessary (as long as it is implied in the context): it would be fine to say *To improve law enforcement, we must sort the good from the bad*. Thus, in our view, (2c) is a valid instance of <*the* A>, while (2b) is not.

Manual exclusion of false positives left 1,782 valid tokens.

4. <D $N_{pl}$ $N_{sing}$>

The automatic process found cases of **at nn2 nn1**, where neither of the nouns is capitalized and the first noun does not contain a hyphen. This yielded 817 tokens.

The most common plural noun by far was *police*, which occurred in 178 tokens (marked *p*), e.g. *a police officer*. In our hand-filtering, we excluded these, simply because we felt phrases with this word might have too much effect on the result. (We also ran the tests reported in the paper with the inclusion of phrases with *police*, and the results were almost unchanged: the observed-to-expected vc-to-sc ratio was 2.247 when *police* was excluded, and 2.246 when it was included.)

False positives were of several kinds: (i) the preterminals were mislabeled; (ii) the identified phrase combined parts of two NPs, e.g. *all the things society has said a woman's home should be*; (iii) the identified head noun (the third word of the phrase) was actually a noun modifier of a following noun, e.g. *the standards adoption process* (these cases are marked *l*); (iv) (most common) the apparent plural noun was actually a possessive noun in which the apostrophe had been omitted, e.g. *the meeting might add to the employees stress*. Most of the borderline cases involved the last of these possibilities. An example is *the kids area at McDonald's*. It seems possible to interpret this phrase as it is, but adding an apostrophe (*the kids' area*) seems more standard; thus, this token was excluded.

Manual exclusion of false positives left 307 valid tokens.

5. <$N_{sing}$ A $N_{sing}$>

The automatic process found tokens of **nn1 jj nn1**, not preceded by [**adjmn**] or followed by **n**. Many false positives arose due to mislabeled words (e.g. *surgeons will practice virtual surgery*) or when the word sequence overlapped two constituents (e.g. *The attempted military coup … is in reality open aggression*). In other cases, the first noun and adjective really formed an adjectival expression (with the usual hyphen omitted), e.g. *self destructive* or *state sponsored*. We marked such cases as *h* and did not consider them valid tokens. In some cases, capitalized phrases were allowed, e.g. *World Scientific Press* (after all, this was a choice on someone's part to create an <$N_{\text{sing}}$ A $N_{\text{sing}}$> phrase), but not if one of the words was a proper name (e.g. *Wexner Medical Center*).

Sometimes it was difficult to decide whether the second word was a true adjective or the first word of a two-word noun. A phrase like *high school* seems like a noun, for all intents and purposes; the same with *black magic* and *big shot*. One might also describe these as idiomatic expressions; their meaning is somewhat arbitrary and conventional and cannot be inferred simply from knowledge of the word meanings. Thus, the phrases *Heritage High School* and *TV big shot* were excluded. This is in contrast to phrases like *opposition political activity*, *state legislative intent*, *classroom interactive video*, *student demographic information*, or *quality financial advice*; while the second and third words of these phrases may form common expressions, they can be understood without having been encountered before. There were, however, many judgment calls in this regard. To decide such cases, we used the Merriam-Webster online dictionary. This lists e.g. *high school* and *working group* as nouns, but not *social worker* or *medical care*. Clearly there is a continuum in this regard between two-word nouns and productive adjective-noun expressions, and imposing a hard cutoff between them seems quite arbitrary. For this reason, the identification process for this rare expansion seems especially debatable and subjective.

Manual exclusion of false positives left 69 valid tokens.

## Appendix B. Publicly released data and scripts

The public release of our data and scripts (available at http://davidtemperley.com/inverse-frequency) consists of the following:

1. *Perl scripts used for the automatic search processes.* There is a script for each rare expansion, as indicated below.

| | |
|---|---|
| <$N_{\text{sing}}$ $N_{\text{prop}}$ $N_{\text{prop}}$> | propn-process.pl |
| <*a* $N_{\text{prop}}$ $N_{\text{prop}}$> | propa-process.pl |
| <*the* A> | detadj-process.pl |
| <D $N_{\text{pl}}$ $N_{\text{sing}}$> | plnm-process.pl |
| <$N_{\text{sing}}$ A $N_{\text{sing}}$> | nadjn-process.pl |

The first command-line argument indicates the input file, which must be in the format specified in Section I of Appendix A (*textblock# word lemma preterminal*).
For all common expansions, we used the script common-process.pl. In this case, the first command-line argument indicates which common expansion is to be searched for, using the numbering of common expansions shown in Section II of Appendix A. The second command-line argument specifies the input file. For example, "./common-process.pl 1 [input-file]" will search for <D $N_{\text{sing}}$>.

For all scripts, the output shows the line number in the input file of the first word of the expansion, followed by the 20-word context of the target phrase (which may or may not be a valid token of the expansion); the phrase is marked with square brackets.

2. *Output files of the automatic search processes, along with hand-edits indicating valid tokens*, for the rare expansions only. Script names correspond to expansions as shown above; for example, propn.txt corresponds to <$\mathrm{N}_{\mathrm{sing}}$ $\mathrm{N}_{\mathrm{prop}}$ $\mathrm{N}_{\mathrm{prop}}$>. The files show, for each target phrase, (i) the line number in the input file on which the first word of the expansion occurs, (ii) a letter indicating its validity (*g* = valid; *x* or anything else indicates invalid, as explained in Section III of Appendix A), (iii) the 20-word context around the target phrase; the phrase is marked with square brackets.
3. *A script for processing the output files*. The script cluster.pl takes one of the 15 output files described above, either hand-annotated or not. (For hand-annotated files, it takes only tokens marked *g* as valid; for unannotated files, it takes all tokens to be valid.) It finds the statistics used in our analysis: (i) the number of very close and somewhat close repetitions, (ii) the ratio between those counts, (iii) the same statistics for a random distribution of the same number of tokens, (iv) the ratio between the observed and expected "very-close-to-somewhat-close" ratios, and (v) the number of tokens used in coordinate expressions (*A and B*). The first command-line argument is the number of words in the expansion (2 or 3); the second argument is the desired level of verbosity (the amount of information displayed: 0, 1, or 2); the third argument is the file to be processed.

## Address for correspondence

David Temperley
Eastman School of Music
University of Rochester
26 Gibbs St.
Rochester, NY 14604
USA

dtemperley@esm.rochester.edu
https://orcid.org/0000-0002-0569-7877

## Publication history

Date received: 20 July 2022
Date accepted: 19 February 2023
Published online: 9 October 2023