COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2015) 1–22

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12215

Information Density and Syntactic Repetition

David Temperley,[a] Daniel Gildea[b]

[a]*Eastman School of Music, University of Rochester*
[b]*Computer Science Department, University of Rochester*

Received 6 November 2013; received in revised form 23 September 2014; accepted 25 September 2014

Abstract

In noun phrase (NP) coordinate constructions (e.g., NP *and* NP), there is a strong tendency for the syntactic structure of the second conjunct to match that of the first; the second conjunct in such constructions is therefore low in syntactic information. The theory of uniform information density predicts that low-information syntactic constructions will be counterbalanced by high information in other aspects of that part of the sentence, and high-information constructions will be counterbalanced by other low-information components. Three predictions follow: (a) lexical probabilities (measured by N-gram probabilities and head-dependent probabilities) will be lower in second conjuncts than first conjuncts; (b) lexical probabilities will be lower in matching second conjuncts (those whose syntactic expansions match the first conjunct) than nonmatching ones; and (c) syntactic repetition should be especially common for low-frequency NP expansions. Corpus analysis provides support for all three of these predictions.

Keywords: Information; Syntax; Coordination; Language production; Probabilistic models

1. Introduction

Information, in the technical sense of the term, is the negative log of probability: Less probable events convey more information (Shannon, 1948). Informally speaking, the information of an item in a sequence of items is a measure of how unexpected or surprising it is. In recent years, the concept of information has attracted considerable interest in language research. It has been shown that the information of a word in a sentence (sometimes known as its *surprisal*) is a highly effective predictor of its comprehension difficulty, as reflected in reading time and other measures (Hale, 2001; Levy, 2008). It has also been proposed that communication is optimal when the amount of information conveyed per unit time maintains a fairly consistent, moderate level, close to but not above the "channel capacity" of the perceiver; this is the theory of *uniform information density* (henceforth UID) (Levy & Jaeger, 2007).

Correspondence should sent to David Temperley, Eastman School of Music, University of Rochester, Rochester, NY 14604. E-mail: dtemperley@esm.rochester.edu

The UID theory has been put forth to explain a variety of linguistic phenomena. The theory predicts that lower-probability elements should be more extended in time, and this has been confirmed; for example, words and syllables that are less predictable in context tend to be pronounced more slowly (Aylett & Turk, 2004; Bell et al., 2003), longer words in written text tend to be lower in contextual probability (Piantadosi, Tily, & Gibson, 2011), and contracted forms of verbs (e.g., *I've, he's*) are more likely to be used when the verbs are low in information (Frank & Jaeger, 2008). In addition, it has been shown that sentences occurring later in a discourse tend to contain less frequent words and word combinations (Genzel & Charniak, 2002); this too is explicable in terms of UID, on the grounds that the preceding sentences in a discourse make later sentences more predictable (lower in information), allowing them to have higher information content in other respects. UID has also been applied to more specific phenomena: Jaeger (2010) shows that, in sentences with complement clauses, the optional complementizer *that* tends to be used more often when a complement clause is low in probability given the preceding verb; the complementizer allows the information of such word sequences to be more spread out over time, softening the "spike" of information that would otherwise occur.

The current study applies the UID theory to the domain of syntax. It is widely known that syntactic constructions vary greatly in probability, and that sentence processing is sensitive to these distinctions (Jurafsky, 1996; Levy, 2008; Trueswell & Tanenhaus, 1994). The UID theory predicts that, if a particular syntactic construction is low in probability in a certain context, its high information should be balanced by low information in other aspects of the sentence—for example, in lexical choices or in neighboring parts of the syntactic structure. This point can be seen most clearly in a probabilistic context-free grammar, in which the probability of a syntactic expansion depends only on its parent type and the probability of a word depends only on its syntactic category; in such a situation, the total information for a given region of the sentence will be the summed log probabilities of all the expansions, and high information in some expansions should trade off with low information in others. By all accounts, language is *not* context-free, either in production or in comprehension, and involves many kinds of complex dependencies between the elements of a sentence; for example, the probability of a syntactic expansion depends not only on its parent but also on its grandparent and more distant ancestors (Johnson, 1998; Klein & Manning, 2003), the probability of a particular verb depends on the subject noun (Trueswell & Tanenhaus, 1994), and the probability of a noun taking a certain syntactic role (such as subject versus object) depends on its animacy and other properties (Traxler, Morris, & Seely, 2002). The probabilities of these components are also affected by factors external to the sentence that are very difficult to quantify, such as pragmatic and discourse factors (Crain & Steedman, 1985; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Because of these complications, any attempt to construct the informational contour of a sentence can only be approximate. Still, if informational tradeoffs between the components of a sentence are observed, and no other explanation for them seems satisfactory, they would seem to be plausibly explained by UID and to offer further support for the theory.

An important factor affecting syntactic probabilities is repetition: The occurrence of a syntactic construction increases its probability of occurring again. Repetition has been observed with a variety of syntactic constructions across a wide range of situations and paradigms. Perhaps the most widely studied phenomenon of this type is syntactic priming: When someone reads (aloud or silently) or hears a syntactic construction in a sentence, such as the passive construction, he or she is more likely to use it in a subsequent sentence (Bock, 1986; Pickering & Ferreira, 2008). More relevant to the current study, repetition effects have also been observed within sentences. Dubey, Keller, and Sturt (2008) found that the occurrence of a particular expansion of a noun phrase (such as NP + prepositional phrase, or NP + complement clause) in a sentence increases the probability that it will recur in the same sentence. In a study of syntactic repetition in speech, Reitter, Keller, and Moore (2011) observed a general tendency toward repetition, peaking at the minimum distance of one second and decreasing steadily as a function of time; while they did not distinguish within-sentence and between-sentence repetitions, the frequency of repetitions at very short time intervals suggests a high incidence of within-sentence repetition. (In some cases syntactic repetition can serve a grammatical function, such as the correlative comparative in English, e.g., *the more the merrier*, but such phenomena are not our concern here.)

The tendency toward repetition has been found to be especially strong within coordinate structures (in this case it is sometimes known as *syntactic parallelism*). For example, in a phrase of the form NP *and* NP, there is a high probability that the two conjoined NPs will be similar in syntactic structure. In the abovementioned study, Dubey et al. (2008) found that the general tendency toward repetition of NP expansions was significantly greater for conjoined NPs in a coordinate phrase. (Further evidence for this phenomenon will be presented below.) This effect also appears to influence comprehension. In a study by Frazier, Munn, and Clifton (2000), subjects read sentences containing conjoined NPs with the same syntactic structure, such as (1a), or different structures, such as (1b):

(1a) Hilda noticed *a strange man* and *a tall woman* when she entered the house.
(1b) Hilda noticed *a man* and *a tall woman* when she entered the house.

Subjects read *a tall woman* more quickly in the first sentence than in the second, suggesting they expected the syntactic structure of the first NP to be repeated in the second. Notably, no such effect was observed when the two NPs occurred in the same sentence but were not conjoined.

In a coordinate construction in which the syntactic structure of the second NP matches that of the first, the syntactic structure of the second NP is relatively high in probability (given the context) and thus low in information. We will say that the probability of the second NP receives a *repetition boost*; by contrast, the first NP receives no such boost. From the perspective of UID, we would expect the low syntactic information of the second NP to be balanced by high information in other components in this part of the sentence—specifically, with regard to lexical probabilities and other syntactic choices. Consider the sentence below (from the Brown corpus, a corpus drawn from a variety of sources of written English):

(2) A tug-of-war between *an old bottle* and *an inefficient corkscrew* may do as much harm as a week at sea.

(The point of the sentence is that old wine can be damaged if the bottle is shaken.) For now, let us define the lexical probabilities of a phrase crudely as the product of the individual probabilities (frequencies) of the words taken out of context (more context-dependent measures will be considered below). It seems intuitively clear that the lexical probabilities of the second NP are lower than those of the first by this measure (and corpus data confirms this): *Inefficient* is less common than *old*, and *corkscrew* is less common than *bottle*. The high syntactic probability of the second conjunct (high because it syntactically matches the first conjunct) is therefore counterbalanced by low lexical probabilities. UID predicts that the sentence above would be more typical than the same sentence with the two conjuncts reversed:

(3) A tug-of-war between *an inefficient corkscrew* and *an old bottle* may do as much harm as a week at sea.

In (3), the information flow of the sentence is highly unbalanced, as both syntactic and lexical probabilities are lower in the first conjunct.

Other theories also make predictions about the ordering of conjuncts that may be difficult to disentangle from the predictions of UID: notably, that "given" elements in a discourse tend to precede "new" ones (Arnold, Wasow, Losongco, & Ginstrom, 2000; Clark & Haviland, 1977). UID makes a further prediction, however, that does not appear to be generated by any other theory. A second conjunct will only receive a repetition boost when the first conjunct has the same syntactic structure; thus we predict that a second conjunct with low lexical probabilities such as *an inefficient corkscrew* will more likely occur when the first conjunct matches it syntactically (we will call these *matching* coordinate phrases) than when it does not (*nonmatching* phrases). Thus (2) above should be more typical than sentence (4), in which the first and second conjuncts are syntactically different:

(4) A tug-of-war between *a bottle that is several years old* and *an inefficient corkscrew* may do as much harm as a week at sea.

UID makes one further prediction about coordinate constructions. Consider sentences (5) and (6), from the *Wall Street Journal*:

(5) Its ambiguity and uneasy mixture of *the serious* and *the comic* is no doubt one reason why it is very much in vogue with directors just now.
(6) In fact, there's only one person involved who's happy, and that's Floyd String, *president of Coin Wrap* and *conceiver of the cement-truck solution*.

Like sentence (2) above, both of these sentences contain matching coordinate structures. In these two cases, however, unlike in sentence (2), the syntactic structure of the two NPs is itself low in probability. In sentence (5), the NP expansion consists of just Det Adj (a determiner plus an adjective)—a very rare expansion of NP, less than 1/30 as

common as Det Adj N. Similarly in (6), the NP expansion is a bare nominal predicate—a singular countable noun with no determiner—also a fairly rare construction (though not uncommon with nouns like *president*).

A rare syntactic construction creates a spike in information that may, in combination with other probabilities, approach or exceed the channel capacity of the perceiver. The UID theory predicts that, when rare syntactic constructions are used, they will tend to be used in situations where some contextual factor increases their probability. The repetition boost provides just such a factor, perhaps boosting the probability of the second NP expansion to a more satisfactory level. (The first conjunct—which contains the same rare syntactic construction—receives no repetition boost; but it may be that the processing of this first conjunct "spills over" to the second conjunct to some extent, so that high information in the first conjunct can be mitigated by low information in the second. We will return to this point.) With a more common expansion such as Det Adj N, the probability of the expansion is already fairly high so it may be that no repetition boost is necessary. Thus, the UID theory predicts that rare syntactic constructions will be (in relation to their overall probability) especially common in matching coordinate constructions.

As mentioned earlier, Reitter et al. (2011) found a general tendency for repetition of syntactic structures. They explain this tendency as an effect of syntactic priming: the general tendency for recently heard or used syntactic constructions to be repeated. While early priming studies (Bock, 1986; Pickering & Branigan, 1998) used experimentally controlled input, more recent studies have observed priming effects in natural corpus data (Gries, 2005; Szmrecsanyi, 2005); these studies also show that a speaker's production patterns can be primed by their own previous production as well as by other input. It has also been found that low-frequency constructions prime more strongly than high-frequency ones—the so-called inverse frequency effect (Hartsuiker & Westenberg, 2000; Scheepers, 2003). Reitter et al. (2011) found that the tendency for syntactic repetition was more marked for less frequent structures, further supporting priming as an explanation for their findings (see also Jaeger & Snider, 2013). Thus, with regard to parallelism in coordinate structures, priming could be said to make the same prediction as UID: that rare syntactic constructions will have an especially strong tendency to occur in matching constructions. Dubey et al. (2008) also invoke priming to explain repetition in coordinate structures, though they do not consider the inverse frequency effect. After demonstrating the empirical validity of the inverse frequency effect for syntactic repetition in coordinate structures, we will argue that it is more convincingly explained by UID than by priming.

To summarize, we have put forth three predictions:

1. In matching coordinate constructions, lexical probabilities will be lower in second conjuncts than in first conjuncts.
2. In second conjuncts, lexical probabilities will be lower in matching constructions than in nonmatching ones.
3. Rare NP expansions will more often be used (in relation to their overall frequency) in matching coordinate structures than common ones.

In what follows, we use corpus methods to test these three predictions. As noted earlier, both the first and third predictions may be predicted by other theories—the first by

discourse salience accounts, and the third by syntactic priming. Therefore prediction 2 is of particular interest, as it is not predicted by any other theory.

Predictions 1 and 2 require a way of defining the lexical probabilities of a sequence of words; how should this be done? One widely used way of defining the probability of a word sequence is with n-gram models (Genzel & Charniak, 2002; Levy, 2008; Piantadosi et al., 2011). Under this approach, the probability of each word is defined by its observed probability in a corpus, conditional on the previous N words; the probability of the entire sequence is defined by the product of these conditional probabilities. Trigram models, in which N = 2, are especially widely used, and we will use them here. (Higher values of N are problematic because many n-grams observed in testing will never have been seen in training.) Following Levy (2008) and others, we do not assume that humans actually compute n-gram probabilities (though they may); n-grams could be viewed instead as an indirect way of capturing a combination of other probabilities that are directly computed (or reflected) in processing, such as unigram (single-word) probabilities, collocations or idioms of arbitrary length, and head-dependent probabilities. Importantly, n-gram probabilities are also affected by syntax: No doubt, word sequences built on uncommon syntactic constructions will generally be low in n-gram probability. But in the tests below, we only compare NPs whose top-level syntactic expansion is the same, thus excluding the influence of this factor. N-gram probabilities may also be affected by lower syntactic expansions within the NPs, and these may differ between the NPs being compared; for example, if two matching NPs expand to NP PP, the probabilities of the expansions within the lower NP may differ between them. But this is desirable: The possibility that the probability of the top-level expansion would trade off with other syntactic probabilities is entirely compatible with the theory.

In measuring n-gram probabilities of a conjunct phrase, we consider only the words within the phrase. It is desirable also to capture the dependencies between the conjunct phrases and the external context. One way of doing this is by examining the dependency between the head of each conjunct phrase and the external head. Here, we assume a "multi-head" view of coordinate phrases (Temperley, 2005) in which the head of each conjunct acts as a dependent of the external head. For example, in the phrase *a tug-of-war between an old bottle and an inefficient corkscrew*, both *bottle* and *corkscrew* are dependents of the preposition *between*. Using corpus frequencies, we can estimate the probability of each conjunct head given the external head. (Here we only consider cases where the external head precedes the coordinate phrase.) The predictions are that these probabilities will be lower for second conjuncts than for first conjuncts, and that they will be lower for matching second conjuncts than for nonmatching ones.

One question that arises here is whether to examine written or spoken data. One might suppose that phenomena of UID would be more evident in speech, where the flow of information is controlled by the producer and could potentially exceed the listener's channel capacity; with written text, the reader can control the flow of information by modulating their reading speed. Several previous studies have found UID effects in written text, however (Genzel & Charniak, 2002; Piantadosi et al., 2011). It is possible that habits

from speech are carried over into writing, or that uneven information flow is disruptive to reading. In addition, the effects under investigation depend on low-frequency words and word combinations, and these are more plentiful in written than in spoken text (Nation, 2006); we suspect that there is also more syntactic variety in written text, though this has not been proven. (Rare constructions such as those in (5) and (6) above seem more characteristic of written text.) Thus, we predicted that the effects at issue would emerge more strongly in written than in spoken text. In the tests that follow, our main focus is on written data, though we briefly consider speech data as well.

2. Quantifying the repetition boost

Before proceeding, we must consider one preliminary issue. An essential premise of our study is that matching coordinate constructions are high in probability—that is, that the syntactic expansion of the second conjunct in a coordinate NP is more likely than chance to match the expansion of the first conjunct. There is suggestive evidence for this; in particular, Dubey et al. (2008) examined five common NP expansions—N, Det N, Det Adj N, NP PP, and NP SBAR—in coordinate structures and found a tendency toward repetition in all five cases (see also Levy, 2004). But the amount of repetition across all NP expansions has never been examined. It is also of interest to quantify the difference in probability between a matching second conjunct and a nonmatching one—what we earlier called the "repetition boost."

Our primary source of test data, in this test and throughout the study, is the Penn Treebank *Wall Street Journal* corpus (hereafter the WSJ corpus), about 1 million words of text from the *Wall Street Journal* from 1987 to 1989, hand-annotated with parts of speech and syntactic constituents (Marcus et al., 1994). The example below illustrates the Penn Treebank conventions for the labeling of two-conjunct coordinate NPs:

(7) (NP (NP (DT an) (JJ old) (NN bottle)) (CC and) (NP (DT an) (JJ inefficient) (NN corkscrew)))

The outer NP contains the whole coordinate phrase; the two conjuncts are also labeled as NPs. CC is the coordinating conjunction, which is *and* in 92.4% of tokens and *or* in 6.5%; in the remaining 1.1% of tokens it is *but*, *plus*, *versus*, or *nor*. Each word is assigned its own preterminal (part-of-speech category): DT for determiner, JJ for adjective, NN for singular noun, and so on. A different labeling convention is used for so-called coordinate NPs in which each conjunct is only a single word (sometimes known as "binomial" NPs); in such cases the conjuncts are not given their own NPs but are simply labeled with preterminals, for example, (NP (NN oil) (CC and) (NN water)). We exclude such cases here for the sake of simplicity (and also for another reason, discussed below). We also exclude coordinate phrases with more than two conjuncts. With these restrictions, the WSJ corpus contains a total of 4,956 NP coordinate constructions.

Our focus is on the syntactic expansions of the conjunct NPs. Conjunct NPs (like NPs in general) can expand in a wide variety of ways. Sometimes they expand to

preterminals, as in (7) above; in other cases they expand to nonterminal constituents, such as NP PP or NP SBAR (seen in relative clause constructions); in other cases they expand to a combination of preterminals and nonterminals, for example, DT NN SBAR, seen in noun complement constructions (e.g., *the fact that...*) (In identifying the expansion of an NP, we ignore "empty categories," i.e., constituents containing no lexical items.) The question is, given the expansion of the first conjunct, what is the probability of a matching second conjunct versus a nonmatching one. There are 1,005 matching coordinate phrases in the corpus and 3,951 nonmatching ones, showing that nonmatching phrases are more common overall; however, the nonmatching probability mass is divided up among many possible expansions. We want to know the probability of a *specific* nonmatching expansion versus a matching one. (As an example, the expansion DT NN occurs 172 times in first conjuncts. DT NN is also the expansion of the second conjunct in 44 of these 172 cases, more than any other expansion.) We measured the overall tendency toward repetition using the concept of surprisal. Given a first conjunct expansion (NP1), the surprisal of a matching second conjunct expansion (NP2) was calculated as

$$\sum_{\substack{\text{NP1,NP2:}\\ \text{NP2=NP1}}} P(\text{NP1}, \text{NP2}) \log_2 P(\text{NP2} \mid \text{NP1})$$

where P(NP1,NP2) is normalized to sum to 1 over all coordinates in which NP2 = NP1. We also calculated the same quantity for nonmatching coordinates, assuming NP2 $\neq$ NP1. The resulting values are 2.11 for matching conjuncts and 5.70 for nonmatching conjuncts. (The overall surprisal of NP2 given NP1—which is the same as the conditional entropy of NP2 given NP1—is a weighted average of these two quantities, 4.97.) This shows that, indeed, matching expansions are much more probable in context than nonmatching ones. In terms of information theory, a nonmatching expansion carries 3.59 bits more information than a matching one; this is the repetition boost that, according to the UID theory, should be counterbalanced by higher information in other aspects of the second conjunct.

One might wonder why the tendency toward parallelism in coordinate phrases exists at all. To some extent, it may be an artifact of context. By necessity, two NPs in a coordinate structure are similar, indeed usually identical, in their syntactic role: If one is a direct object, the other one is, too. And this may indirectly give rise to similarities between them, as NPs in different roles show different syntactic tendencies (for example, a subject NP is much more likely to be a pronoun than a direct object NP). There may also be semantic factors at work: If one NP in a coordinate structure is a numerical expression, for example, the other one may tend to be as well (perhaps they are both objects of a verb that tends to take numerical expressions as objects). Priming may also play a role, as has been argued by Dubey et al. (2008). The reasons for parallelism as a general phenomenon would be interesting to explore further, but this is not our main concern here.

3. Lexical probabilities in first versus second conjuncts

Our first prediction is that second conjuncts in matching NP coordinate phrases will have lower lexical probabilities than first conjuncts. Differences between first and second conjuncts have been explored in several previous studies. Levy (2004) and Temperley (2005) found a preference for short–long ordering in NP coordinate phrases; Levy also found a preference for "given" before "new" discourse elements independent of length. Also relevant, several studies have explored the ordering of "binomial" phrases—phrases of the form *word1 and word2*, where the two words may be nouns or other categories such as verbs or adjectives. Fenk-Oczlon (1989) found a large frequency effect in the ordering of binomials, with the more frequent item tending to come first; Benor and Levy (2006) found a similar effect, and also examined the role of a number of other factors in the ordering of binomials, several of which will be considered below. It is important to note that the current study excludes binomials; as noted earlier, in coordinate phrases in which the two conjuncts are single words, the conjuncts are not labeled as NPs and are therefore not part of our data set.

We compared the lexical information of first and second conjunct phrases in the following way. Using the set of NP coordinate phrases in the WSJ corpus, we extracted just the 1,005 matching phrases. We created a list of expansions that occurred in at least one token; 88 such expansion types were found. (Since the phrases were matching, both NPs had the same expansion.) We then calculated what we will call the *internal lexical information* of each phrase, by computing the product of the conditional n-gram probabilities of the words. Our n-gram model combined unigram, bigram, and trigram probabilities, using Kneser–Ney smoothing (Kneser & Ney, 1995); for the first and second words of each NP we used unigram and bigram models, respectively. (N-gram probabilities were defined using a corpus of 43 million words from the *Wall Street Journal*.[1]) We then took the negative log (base 2) of the lexical information of the phrase, and divided this by the number of words. This could also be viewed as the log of the perplexity of the phrase. For each expansion, we found the average information of all first conjuncts with that expansion, and then did the same for second conjuncts. As predicted, the mean information across expansions was higher for second conjuncts (10.05) than for first conjuncts (9.49); a paired t-test across expansions showed that this was highly significant, $t(87) = -2.54$, $p < .01$. The number of syntactic expansions showing the expected pattern, 54 of 88, was significantly greater than half ($\chi^2(1) = 4.6$, $p < .05$), showing that the effect occurs quite broadly and is not limited to a few syntactic constructions.

As discussed earlier, we also measured lexical information in a more contextual way, based on the probability of the head of the conjunct phrase given the external head. In this case, we only considered tokens in which the external head preceded the conjunct phrase, since it is presumably only in these cases that the external head could be used to predict (lower the information of) the conjunct phrase; we call these "right-branching" coordinate phrases. An algorithm by Collins (1999) was used for extracting head-dependent relationships from Treebank-style parsed text. Statistics regarding the probability of

each conjunct head given an external head were gathered from the WSJ training data (using head-modifier relations from automatically generated parses). We used Kneser–Ney smoothing to estimate probabilities for unseen head-modifier pairs. We then used the same procedure as in the previous test. In this case, we only considered expansions that occurred in at least one right-branching matching coordinate phrase; 80 such expansions were found. Averaging across expansions, the average information (negative log probability) of conjunct heads given external heads was 14.40 for first conjuncts, 14.94 for second conjuncts; this difference in means was in the predicted direction but not significant, $t(79) = -1.17$, $p = .12$. However, the number of expansions showing the predicted pattern, 51 of 80, was significantly greater than half ($\chi^2(1) = 5.5$, $p < .05$). We also examined the probabilities of the heads themselves (i.e., their overall probabilities out of context); again, the difference between information across expansions was in the predicted direction but not significant ($M_{\text{first}} = 13.87$, $M_{\text{second}} = 14.48$, $t(79) = 1.22$, $p = .11$); the proportion of expansions showing the predicted pattern, 50 of 80, was significantly more than half ($\chi^2(1) = 4.5$, $p < .05$).

The tests presented above show that first conjuncts are significantly higher in internal lexical information than second conjuncts. Regarding the probabilities of conjunct heads, and of heads given the external head, the difference between first and second conjuncts is still evident, though weaker (the proportion of expansions showing the pattern is significant, but the difference in information between first and second conjuncts is not). While these findings are in accord with the UID theory, there are several possible explanations for them besides UID. First of all, there is a large difference in length between first and second conjuncts. Over all NP CC NP constructions in the WSJ corpus, the mean length of the first NP is 3.28 words while that of the second is 4.39. It seemed possible that longer phrases would tend to have higher perplexity (though it is not obvious why this would occur), which might then account for the observed difference. To eliminate this possible confound, we recalculated internal lexical information considering only cases in which the two NPs were the same length in number of words: 621 such tokens were found. (This yielded a somewhat smaller pool of expansions: 73 instead of 88.) Once again, second conjuncts had significantly higher internal information ($M_{\text{first}} = 10.09$, $M_{\text{second}} = 10.62$, $t(72) = 2.18$, $p < .05$), suggesting that the difference in length does not explain the difference in information content. Out of 73 expansions, 45 (61.6%) showed the predicted effect, virtually the same proportion as was found when all conjuncts were considered (54/88 = 61.4%); due to the smaller number of data points, however, the difference from an even split in this case was only marginally significant ($\chi^2(1) = 3.96$, $p = .06$).

A more difficult problem concerns semantic differences between first and second conjuncts. Levy (2004) observes that second conjuncts are more likely than first conjuncts to represent new rather than given discourse entities; this is not surprising, in light of the general tendency for given discourse elements to precede new ones (Arnold et al., 2000; Clark & Haviland, 1977). Related to this, Benor and Levy (2006) find that the second element in a binomial phrase tends to be both "formally marked" and "perceptually marked"; they acknowledge that both formal markedness and perceptual markedness are

closely related to the given/new distinction, with marked elements more likely to be new. (Benor and Levy use the same corpus used here, but since they consider only binomials, their data set does not overlap with ours.) The given/new distinction is difficult to encode in a rigorous way. It is generally agreed that given entities may be only implied by the context, not stated explicitly (Arnold et al., 2000; Prince, 1981). For example, in a discussion of a corporation, the president of the corporation might be taken as a given entity (in that its existence is assumed) even if not explicitly mentioned. A further point is that the tendency toward given-before-new ordering in coordinate constructions might, in itself, be explained as an effect of UID: Given discourse entities are presumably more predictable than new ones, so we would expect them to occur in higher information syntactic environments, that is, in first conjuncts rather than second conjuncts. Still, it would be of interest to try to tease apart the effects of information content and discourse newness on the ordering of conjuncts. We will not attempt that here but leave it as a project for the future.

A further factor in the ordering of conjuncts is what has been called *iconicity*: When there is an implied temporal or causal ordering between the two conjuncts, this tends to be reflected in their syntactic ordering. This has been observed as a factor in binomials, reflected in common phrases such as *birth and death* and *parent and child* (Benor & Levy, 2006; Fenk-Oczlon, 1989). No doubt it is a factor in phrasal coordinate constructions as well. For example, in sentence (2) above, restated here:

(8) A tug-of-war between an old bottle and an inefficient corkscrew may do as much harm as a week at sea.

the causal priority of bottle over corkscrew (one only needs a corkscrew if one has a bottle) may well be a factor in the ordering of the conjuncts. It is not obvious that temporally or causally prior elements would tend to be lower in lexical information content, but it is certainly a possibility.

Because of these semantic confounds, it must be admitted that our comparison of first and second conjuncts is not wholly convincing as a test of the effect of UID on coordinate constructions. The test we present next is much less vulnerable to this criticism, and thus offers a more persuasive test of the role of information flow in syntax.

4. Lexical probabilities in matching versus nonmatching second conjuncts

Our second prediction is that lexical information will be higher in matching second conjuncts than in nonmatching ones. We tested this using a similar procedure to that used to compare first and second conjuncts. We used the same data set of coordinate phrases in the WSJ corpus. In this case, we included only expansions that occurred in at least one matching token and one nonmatching token; this yielded 73 expansion types. The internal lexical information of each phrase was defined in the same way as in the previous section. For each expansion, we found the average log perplexity of all matching second conjuncts with that expansion, and then did the same for nonmatching second conjuncts

with the same expansion. As predicted, the mean information across expansion types was significantly higher for matching second conjuncts than for nonmatching ones ($M_{\text{matching}} = 10.31$, $M_{\text{nonmatching}} = 9.79$, $t(72) = 2.13$, $p < .05$). The number of expansions showing the predicted pattern, 49 of 73, was significantly greater than half ($\chi^2(1) = 8.6$, $p < .005$).

We also compared matching and nonmatching second conjuncts with regard to head probabilities and the probabilities of heads given external heads. As before, we exclude the cases where the external head follows the coordinate phrase; this yielded 67 expansions that occurred in at least one matching and one nonmatching token. Across expansions, matching conjunct heads had significantly higher information than nonmatching heads ($M_{\text{matching}} = 14.92$, $M_{\text{nonmatching}} = 13.94$, $t(66) = 1.99$, $p < .05$); the number of expansions showing the predicted pattern, 39 of 67, was more than half but not significantly so ($\chi^2(1) = 1.5$, $p = .22$). For the probabilities of heads given external heads, the difference was in the predicted direction and marginally significant ($M_{\text{matching}} = 14.26$, $M_{\text{nonmatching}} = 13.31$, $t(66) = 1.59$, $p = .06$); the number of expansions showing the predicted pattern, 39 of 67, was more than half but not significantly ($\chi^2(1) = 1.5$, $p = .22$).

Notably, the comparison of matching and nonmatching conjuncts avoids the semantic confounds that arose with the comparison of first and second conjuncts. While first and second conjuncts tend to differ with regard to factors such as discourse newness and temporal/causal priority, it has never been suggested that matching second conjuncts differ from nonmatching ones in these ways, and it is difficult to imagine why they would. Thus, the test of matching and nonmatching second conjuncts provides a more convincing demonstration of the role of UID in coordinate constructions than the comparison of first and second conjuncts.

All of the predictions tested so far have been confirmed more strongly when probabilities are computed over all the words in the phrase (using N-grams), rather than for the conjunct head alone or for the conjunct head given the external head. When N-grams are used, second conjuncts reflect higher information than first conjuncts both in terms of the mean information across expansions and in terms of the proportion of expansions showing the pattern; similarly, matching second conjuncts are higher in information than nonmatching ones by both measures. For head probabilities and head-dependent probabilities, by contrast, differences between first and second conjuncts and between matching and nonmatching conjuncts are significant by only one of the two measures (though all differences are in the predicted direction). We can think of several possible methodological reasons why the N-gram results are stronger than the head and head-dependent results. First, the N-gram tests simply provide more data, because probabilities are being computed for multiple words in each phrase rather than just one; this allows the fairly subtle differences between phrase types to emerge more strongly. (With head-dependent tests, there is an additional problem of sparse training data: About 28% of the head-dependent tokens in the test set were never seen in training, in which case the unigram probability of the conjunct head was used instead.) In addition, the identification of conjunct heads (and external heads) relies on an automatic algorithm for extracting dependencies from constituent trees which may not always be accurate, both in the training data and the test

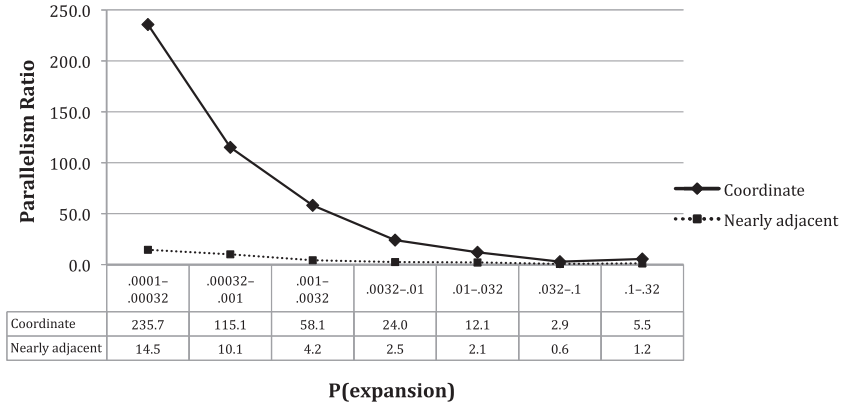
data. (In the training data, the identification of the constituent trees is itself automatic, an additional source of error that affects all the tests reported here.) Replicating these tests with a larger and more accurate training dataset would certainly be desirable.

5. Repetition with rare versus common syntactic expansions

Our third prediction is that the tendency toward repetition in coordinate constructions will be stronger for low-frequency syntactic expansions than for high-frequency ones. To test this prediction, we first calculated the probability of all NP expansions across all NPs. (We used the hand-parsed WSJ corpus as training data for this test because 1 million words of data seemed sufficient and because it seemed possible that the larger automatically parsed corpus might contain systematic errors, especially with rare constructions.) These probabilities represent the probability of an NP expansion (call it X) occurring in the second conjunct, independent of context. Multiplying this by the number of first-conjunct occurrences of X gives an indication of the number of matching conjunct phrases with X that would be expected due to chance—that is, if the expansion of the second conjunct was "context-free" and depended only on the parent. We compared this with the number of matching phrases with X that was actually observed; the ratio between the observed count and the expected count gives an indication of the tendency toward parallelism with that expansion—what we will call the "parallelism ratio."

Across all 4,956 tokens in the data set, there were 1,005 matching tokens, compared to an expected count of 170.6, for a parallelism ratio of 5.89. This gives another measure (in addition to the surprisal-based one presented earlier) of the general tendency toward repetition in coordinate phrases. Of greater interest in the current context, we can also split the data according to the frequency of the first conjunct expansion. If we consider only expansions whose context-free probability is greater than .01 (there are just 20 such expansions, though they account for more than half the tokens), the ratio is 4.84, slightly lower than the value for the entire data set. By contrast, if we consider expansions whose probability is less than .0001, the parallelism ratio is 1,281.39. These findings are suggestive, but they should be interpreted with caution. In the low-probability category, only 10 of the 263 expansions in the category have any matching tokens at all, but because the expected number of matches is so low (much less than 1), the parallelism ratio is still high. One might wonder whether this effect was due to a few idiosyncratic rules, or perhaps to a certain kind of construction that takes several different forms. Inspection of the data made this seem unlikely; the ten expansion types in the low-frequency category reflect a variety of constructions with no apparent commonality between them.

To explore this further, we grouped the rules into probability "bins" with endpoints defined by $10^{x/2}$ for x = (integers) -8 through 0; this creates an overall range from .0001 to 1, divided on a logarithmic scale into 8 sub-ranges. The relationship between probability and parallelism ratio for each bin is shown in Fig. 1. (The highest-probability bin contained no rules and is therefore not shown.) Both ends of the probability range suffer somewhat from sparse data; in the highest-probability bin, there is only one rule within



	.0001–.00032	.00032–.001	.001–.0032	.0032–.01	.01–.032	.032–.1	.1–.32
Coordinate	235.7	115.1	58.1	24.0	12.1	2.9	5.5
Nearly adjacent	14.5	10.1	4.2	2.5	2.1	0.6	1.2

Fig. 1. Relationship between parallelism ratio for NP expansions (ratio of observed frequency to expected frequency in matching constructions) and overall probability of expansion. Expansions are binned by probability; numbers below the horizontal axis show the probability range for each bin.

the probability range (the expansion NP PP); in the lowest-probability bin, there are only seven matching tokens. Still, there is a very clear and consistent effect of increasing parallelism as rule frequency decreases. To test the significance of this effect, we performed a mixed logit regression (Jaeger, 2008) to analyze the occurrence of repetition based on log odds of the frequency of an expansion, while controlling for the identity of the individual expansion. We included an offset for the log odds of the expansion, in order to remove the effect of repetitions that would result by chance if the first and second expansions were independent:

$$\text{logit}(r_i) = \alpha + \beta\text{logit}(p_i) + \text{logit}(p_i) + u_i$$

where r_i is the probability of repetition for expansion i, p_i is the overall probability of expansion i of an NP, and $u_i \sim N(0, \sigma)$ is a random intercept for expansion i. This yielded a parameter estimate $\beta = -0.64$ for the log odds of the expansion, with a Wald's Z of -8.645, significant at $p_Z < 10^{-15}$. This shows that the tendency for repetition is stronger with less frequent expansions.

One aspect of our argument here requires a bit more explanation. With regard to second conjuncts, the idea is that the information "spike" of a rare construction is mitigated when it appears in a matching context, since the syntactic structure of the second conjunct is highly predictable in this case. But of course (assuming a matching context) the same construction also appears in the first conjunct, and here its probability is not increased by the subsequent repetition; this might appear to be an information spike that is not mitigated by any other factor. The UID argument might apply even here, however, due to "spillover" effects: The processing of one segment continues as subsequent segments are encountered (Rayner & Duffy, 1986). If we assume, following Levy (2008), that the processing cost of a word is proportional to (or at least increases with) its information, then it makes sense to regard the information of a word as spilling over to

subsequent words; in that case, the information spike of a high-information word or phrase may be mitigated if it is followed by low-information ones. Presumably, though, the repetition boost has a greater effect on the second conjunct than on the first; indeed, this is crucial for our first two predictions.

An alternative account for our findings is based on the idea of syntactic priming. In a study of five common NP expansion types, Dubey et al. (2008) found a general tendency toward repetition within sentences but found that the tendency was especially strong within coordinate phrases. Dubey et al. attributed this effect to priming. They noted that priming effects decay quite quickly and that the two NPs in a coordinate phrase tend to be especially close (separated just by one word), so it is expected that priming between the NPs of a coordinate phrase would be especially strong. According to this account, the greater tendency for repetition with low-frequency expansions could be attributed to the fact that priming is generally stronger with rare constructions—the so-called inverse frequency effect. According to this view, repetition of NP constructions in coordinates phrases is simply due to the proximity of the two NPs; therefore, it should be observed just as strongly with NPs pairs that are not part of a coordinate constructions but are separated by the same distance, that is to say, with one word between them. To test this, we extracted all pairs of NPs in the WSJ corpus that were separated by one word, excluding those in coordinate constructions; we will call these "nearly-adjacent NP pairs." Such pairs appears in a wide variety of contexts, such as direct-object constructions (NP verb NP) and prepositional phrase constructions (NP preposition NP). The parallelism ratio for this entire data set—the ratio of the number of observed matches between the two NP expansions to the expected number based on the overall frequency of the expansion—is just 1.001, compared to a ratio of 5.89 for the NP-coordinate data set; this indicates that the overall tendency toward repetition for nearly adjacent NPs (when coordinate NPs are excluded) is very slight. We then performed the same test reported above. An effect of more parallelism for less frequent expansions was again found, but it was much weaker: Wald's $Z = -3.001$, significant at $p_Z < .005$. The data are shown in Fig. 1, with expansions binned by overall probability as with the coordinate NPs. The increase in repetition as expansion frequency decreases replicates the inverse frequency effect for syntactic repetition found by Reitter et al. (2011), and it may indicate an effect of priming. But the amount of parallelism for coordinate NPs is far higher than for nearly adjacent NPs, at all frequency levels. Thus, it is clear that syntactic repetition within coordinate NPs cannot be attributed to a general effect of short-distance priming.

In defense of the priming account, one might note that NPs within a coordinate construction occur in similar—indeed, virtually identical—syntactic contexts: Both NPs are within a larger NP and within the same higher level structure as well (for example, a direct object within a relative clause, or whatever it might be). By contrast, nearly adjacent NPs might occur in quite different syntactic environments. It may be that priming effects are stronger when the prime and target constructions occur in similar contexts, in which case the greater priming effect for coordinate NPs would be predicted. However, prior work on priming has cast doubt on this idea. In a study of the dative alternation (*I gave the boy a puppy / I gave a puppy to the boy*) in which the prime and target could

be in either main or subordinate clauses, Branigan, Pickering, McLean, and Stewart (2006) found that priming effects were not significantly affected by whether the prime and target were in the same syntactic environment (both in main clauses or both in subordinate clauses). This suggests that, in general, syntactic context has little effect on priming (Pickering & Ferreira, 2008). To advocate a priming explanation for our results, one would need to explain why this general finding does not apply in the case of coordinate NPs.

6. Syntactic repetition in speech

The tests reported so far have all used written text. As noted earlier, we predicted that the effect of syntactic repetition on lexical information flow would emerge more strongly in written than in spoken text; nevertheless, it seemed wise to examine some speech data as well. For training data, we used a portion of the British National Corpus containing 10 million words of speech (Clear, 1993). For test data, we used the Switchboard corpus, containing about 900,000 words of syntactically annotated telephone conversations (Godfrey, Holliman, & McDaniel, 1992). We identified NP CC NP constructions, excluding those containing disfluencies. We then performed the same tests that were reported above for written data.

The spoken data contained 1,730 coordinate NP phrases, 304 matching and 1,426 nonmatching; there were 60 expansions found with at least one matching token. Mean information (average log perplexity) across these expansions was not significantly different between first conjuncts (10.32) and second conjuncts (9.98), $t(59) = 1.31$, $p = .10$. Regarding the information (negative log probability) of conjunct heads, the difference between first conjuncts (15.83) and second conjuncts (15.80) was not significant, $t(57) = -0.08$, $p = .47$, nor was there a significant difference between first conjuncts (16.28) and second conjuncts (16.26) in the probabilities of conjunct heads given the external head, $t(57) = -0.05$, $p = .48$. Likewise, there were no significant differences between matching and nonmatching conjuncts in regard to average log perplexity ($M_{\text{matching}} = 10.08$, $M_{\text{nonmatching}} = 10.04$, $t(40) = 0.13$, $p = .44$), unigram probabilities of conjunct heads ($M_{\text{matching}} = 16.28$, $M_{\text{nonmatching}} = 16.19$, $t(38) = -0.12$, $p = .45$), or probabilities of conjunct heads given external heads ($M_{\text{matching}} = 16.61$, $M_{\text{nonmatching}} = 16.19$, $t(38) = -0.5$, $p = .31$). Chi-square tests on the proportion of expansions showing the predicted patterns were mostly not significant, except in the case of first versus second conjuncts unigram and head-dependent probabilities, where significantly more than half of expansions showed the predicted effects (for unigrams, $\chi^2(1) = 7.6$, $p < .01$; for head-dependent probabilities, $\chi^2(1) = 9.12$, $p < .005$).

As we did with the written data, we also examined the relationship between the frequency of rules and their tendency to appear in matching coordinate expressions. A logistic regression showed increased parallelism for lower frequency expansions, Wald's $Z = -11.531$, $p_Z < 10^{-15}$. Thus, while the differences in information content between first and second conjuncts and between matching and nonmatching conjuncts found in

written data do not emerge strongly in speech, we do find the same tendency toward greater use of rare expansions in matching coordinate constructions. The differences between speech and writing will be discussed further below.

7. Discussion

Because there is a high probability that the second conjunct in a coordinate NP construction will syntactically match the first, the second conjunct in such matching constructions is low in syntactic information. The theory of UID therefore predicts that matching second conjuncts will be relatively high in information in other ways, in comparison to first conjuncts and also in comparison to second conjuncts in nonmatching constructions. Corpus analyses of written text confirmed these predictions. In terms of internal lexical information, measured with N-gram probabilities, matching second conjuncts were found to be significantly higher in information than first conjuncts and nonmatching second conjuncts. These patterns are also reflected in the probabilities of conjunct heads and the conditional probabilities of conjunct heads given external heads, though somewhat less strongly; these probabilities show some tendency to be lower in second conjuncts than in first conjuncts, and lower in matching second conjuncts than in nonmatching ones. While the higher information level of second conjuncts relative to first conjuncts is predicted by a discourse-salience account, the higher information level of matching conjuncts relative to nonmatching ones is not; this is therefore a strong confirmation of the UID theory.

The UID theory also predicts that rare syntactic expansions will be more often used in matching coordinate constructions (relative to their overall frequency) than common ones, since repetition should be most advantageous with expansions that are otherwise low in probability; this prediction, too, was confirmed. While syntactic parallelism and the higher tendency for parallelism in rare constructions could be predicted by a priming account, a comparison between coordinate NPs and noncoordinate NPs separated by one word showed that parallelism in coordinate NPs is not simply due to a general effect of short-distance priming.

On the whole, the effects that we predicted emerged much less strongly in speech than in writing. Speech data showed much less tendency toward higher lexical information in second conjuncts and matching versus nonmatching conjuncts, though it did show a strong inverse relationship between expansion frequency and use in matching constructions. There are several possible reasons for these observed differences between speech and writing. Written and spoken data are very different in character in many ways. In our corpora, the average sentence length in the written test data (the annotated WSJ corpus) is 23.9 words; the average sentence length in the speech data (the annotated Switchboard corpus) is 8.2 words. The entropy of NP expansions in the written data is 6.55; that of the spoken data is 4.12, suggesting that there is much less syntactic variety in the spoken data than in the written data. It has also been observed that writing generally has more lexical variety than speech (Nation, 2006). These differences suggest that, in many respects, spoken data are less complex than written data and therefore less computationally demanding. If (as will be

argued below) the UID effects proposed here are strategies to facilitate comprehension, it may be that there is simply less need for such strategies in speech. It may also be that there is less time to plan and implement such strategies, given the time pressures of speech production. These differences between speech and writing deserve further study; in the following discussion, however, we focus on our results from written data.

We have tried to make as few assumptions as possible about the factors that influence subjective probabilities in language processing. No doubt, the probability of a particular syntactic construction in a certain context is affected by a variety of semantic and situational as well as purely syntactic factors. This is only a problem for our argument if such factors are confounded in some way with the factors that we examine, so that the information content of second conjuncts versus first conjuncts, or matching second conjuncts versus nonmatching ones, differs for some reason other than UID. We have acknowledged that, indeed, there are some possible confounds in the case of first versus second conjuncts (e.g., the tendency for given items to precede new ones, and for the ordering of items to reflect their temporal or causal ordering), raising some doubt about the UID explanation in this case; in the case of matching versus nonmatching conjuncts, the existence of such confounds seems much less likely.

Earlier we provided a quantitative estimate of the repetition boost—the increase in probability for matching second conjuncts relative to nonmatching ones—as 3.59 bits. Using the perplexity measure proposed above, we could estimate the difference in internal lexical information between matching and nonmatching second conjuncts as $10.31 - 9.79 = 0.52$ bits. Thus, the difference in lexical information between matching and nonmatching conjuncts falls far short of counterbalancing the difference between them due to repetition. This may be due, in part, to imprecision in these estimates, but no doubt it is also due to the fact that the construction of sentences is influenced by many factors other than UID, so that we should not expect information flow to be completely uniform. Another point is that the repetition boost is not necessarily the same for all expansion types—either objectively (in terms of corpus frequencies) or subjectively. Indeed, our claim that the tendency toward repetition is greater for rare constructions suggests that the repetition boost should be higher in such cases. One might wonder whether sentence processing reflects this—that is, whether people are sensitive to the fact that rarer constructions are more likely to repeat and adjust their subjective probabilities accordingly. While this is an interesting question, it does not appear to be a problem for our argument. If the subjective repetition boost is higher for rare constructions, the effect of this would be to raise the subjective probabilities of rare matching conjuncts, bringing them closer to those of frequent matching conjuncts, and thus further evening out the information flow.

The usual explanation for UID effects is that they arise to facilitate comprehension (Jaeger, 2010; Levy & Jaeger, 2007). One might also wonder if they could be explained as benefiting production in some way. In the present case, this seems unlikely. Production-based accounts of syntactic choice phenomena have usually focused on the concept of availability. For example, in complement-clause constructions, the optional complementizer *that* is more likely to be used if the following complement clause subject has

not been used previously in the discourse; this may be because "discourse-new" subjects are less cognitively available and thus require more time to access (Ferreira & Dell, 2000). It is difficult to see how such an explanation could apply to the phenomena at issue here. And in any case, availability effects seem much more likely to occur in speech—where there is pressure for fast, fluent production—than in writing; by contrast, the effects at issue here emerged much more strongly in writing than in speech.

If the effects we observe are strategies to facilitate to comprehension, one might wonder about the specific decision-making processes that bring them about. There are several possibilities. It is possible that the producer forms the two coordinate NPs and then considers their information content in deciding how to order them: If they are syntactically matching, the higher information one tends to be placed in second position. This could explain both the difference in information between first and second conjuncts, and the higher information content of matching versus nonmatching second conjuncts. (In nonmatching contexts, the pressure to place the higher information conjunct last would presumably be absent or at least less strong.) Alternatively, it may be that, in some cases, the two conjuncts are formulated at a semantic or conceptual level and assigned a specific order, and then UID comes into play in the way they are expressed syntactically and lexically; in matching contexts, more freedom is taken with the lexical content of the second conjunct. As well as explaining the differences between first and second conjuncts and between matching and nonmatching conjuncts, this could also accommodate our third prediction: Perhaps, when a rare syntactic expansion has been chosen for the first conjunct, there is an inclination to choose a matching structure for the second conjunct. Yet another possibility is that the first conjunct phrase is be formed before the decision to produce a coordinate phrase is even made. This would be another way of explaining our third prediction: The decision to add a second conjunct might arise, in some cases, as a way of mitigating the information spike caused by a rare syntactic expansion in the first conjunct. Further research is required to tease apart these possibilities.

The focus of our study has been on phenomena of syntactic choice. Our assumption is that there are often multiple ways of expressing a thought—semantically and pragmatically similar if not identical—and that considerations such as UID may affect the choice between them. It is also possible that the current argument could be applied to phenomena of grammar. Consider this sentence from the WSJ corpus:

(9) When the chain stores took over, there was no longer a connection between grower and consumer.

(Since the conjoined noun phrases are single words, they are not labeled as NPs in the WSJ corpus and therefore not part of our data set.) As singular count nouns, *grower* and *consumer* would normally require a determiner (one could not say "there was no longer a connection to consumer"), but their use in a coordinate construction appears to waive this requirement. While explanations for the allowance of bare NPs in coordination have been offered in terms of generative syntax (Heycock & Zamparelli, 2003) and optimality theory (deSwart & Zwarts, 2009), the UID theory offers a simpler explanation. A singular count noun without a determiner generally carries a spike of information that could

impede parsing, but when used in coordination, the repetition boost raises the probability of the bare-NP construction, making it less disruptive. Also of interest in this connection are constructions such as these (also from the WSJ corpus):

(10) The picocassette recorder also helped transform the company's reputation *from follower to leading-edge innovator.*

(11) *Bribe by bribe*, Mr. Sternberg and his co-author, Matthew C. Harrison Jr., lead us along the path Wedtech traveled...

This pattern, labeled the NPN construction by Jackendoff (2008), offers another example of singular count nouns used without a determiner. Here, too, the UID theory may offer an explanation. This case is different from previous ones we have considered, in that the two NPs are not used in coordination, but rather joined by a preposition. But as Reitter et al. (2011) and Dubey et al. (2008) have shown, there is some tendency toward within-sentence repetition of syntactic structures in general, even beyond coordinate structures. Thus, we may assume that any kind of syntactic repetition provides a boost in probability which can soften the information spike caused by a rare construction. Another example of this is the English correlative comparative—for example, "The more I learn, the less I know"—which likewise involves an idiosyncratic construction that is repeated within the sentence. The general principle, applying to both syntactic choice and grammar, is that repetition can license constructions whose rarity might otherwise be disruptive to processing.

Note

1. BLLIP 1987-89 WSJ Corpus Release 1, Linguistic Data Consortium catalog number LDC2000T43.

References

Arnold, J. E., Wasow, T., Losongco, T., & Ginstrom, R. (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, *76*, 28–55.

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.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America*, *113*(2), 1001–1024.

Benor, S. B., & Levy, R. (2006). The chicken or the egg? A probabilistic analysis of english binomials. *Language*, *82*, 233–278.

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

Branigan, H., Pickering, M., McLean, J., & Stewart, A. (2006). The role of global and local syntactic structure in language production: Evidence from syntactic priming. *Language and Cognitive Processes*, *21*, 974–1010.

Clark, H., & Haviland, S. (1977). Comprehension and the given-new contract. In R. Freedle (Ed.), *Discourse production and comprehension* (pp. 1–40). Hillsdale, NJ: Lawrence Erlbaum.

Clear, J. (1993). *The British National Corpus*. Cambridge, MA: MIT Press.

Collins, M. J. (1999). Head-driven statistical models for natural language parsing. Unpublished doctoral dissertation, University of Pennsylvania, Philadelphia.

Crain, S., & Steedman, M. J. (1985). On not being led up the garden path: The use of context by the psychological syntax processor. In D. R. Dowty, L. Karttunen, & A. Zwickyy (Eds.), *Natural language parsing* (pp. 320–358). Cambridge, England: Cambridge University Press.

deSwart, H., & Zwarts, J. (2009). Less form — more meaning: Why bare singular nouns are special. *Lingua*, *119*, 280–295.

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

Fenk-Oczlon, G. (1989). Word frequency and word order in freezes. *Linguistics*, *27*, 517–556.

Ferreira, V., & Dell, G. (2000). Effects of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, *40*, 296–340.

Frank, A., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society (cogsci08)* (pp. 939–944). Washington, DC.

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

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 199–206). Stroudsburg, PA: Assocation for Computational Linguistics.

Godfrey, J., Holliman, E., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech, & Signal Processing (IEEE ICASSP-92)*(pp. 517–520). Piscataway, NJ: IEEE Service Center.

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

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of Naacl-2001* (pp. 159–166). Stroudsburg, PA: Association for Computational Linguistics.

Hartsuiker, R., & Westenberg, C. (2000). Word order priming in written and spoken sentence production. *Cognition*, *75*, B27–B39.

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

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

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVA (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Jaeger, T. F., & Snider, N. (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*, 57–83.

Johnson, M. (1998). PCFG models of linguistic tree representations. *Computational Linguistics*, *24*(4), 613–632.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*, 137–194.

Klein, D., & Manning, C. D. (2003, July). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (pp. 423–430). Sapporo, Japan: Association for Computational Linguistics.

Kneser, R., & Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Vol. *1*, pp. 181–184). Detroit, MI: IEEE.

Levy, R. (2004). *The statistical properties of coordinate noun phrases*. Unpublished manuscript. <http://idiom.ucsd.edu/~rlevy/papers/coordination-handout.pdf>, accessed November 27, 2014.

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 *Advances in neural information processing systems*. Cambridge, MA: MIT Press.

Marcus, M. P., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., et al. (1994). The Penn Treebank: Annotating predicate argument structure. In *ARPA Human Language Technology Workshop* (pp. 114–119). Plainsboro, NJ: Morgan Kaufmann.

Nation, P. (2006). How large a vocabulary is needed for reading and listening?. *Canadian Modern Language Review*, *63*, 59–82.

Piantadosi, S., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences USA*, *108*, 3526–3529.

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

Pickering, M., & Ferreira, V. (2008). Structural priming: A critical review. *Psychological Bulletin*, *134*, 427–459.

Prince, E. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), *Syntax and semantics: 14. radical pragmatics* (pp. 223–255). New York: Academic Press.

Rayner, K., & Duffy, S. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory and Cognition*, *14*, 191–201.

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

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

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. (Continued in following volume)

Szmrecsanyi, B. (2005). Language users as creatures of habit: A corpus-based analysis of persistence in spoken english. *Corpus Linguistics and Linguistic Theory*, *1*, 113–149.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632–1634.

Temperley, D. (2005). The dependency structure of coordinate phrases: A corpus approach. *Journal of Psycholinguistic Research*, *34*, 577–601.

Traxler, M., Morris, R., & Seely, R. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, *47*, 69–90.

Trueswell, J., & Tanenhaus, M. (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). Hillsdale, NJ: Lawrence Erlbaum.