Incremental Generation Drives “Efficient” Language Production

Spencer Caplan*
Department of Linguistics
University of Pennsylvania

November 12, 2018

Abstract

A major testing ground for mechanistic accounts of language production is the study of ‘syntactic optionality’; i.e. given multiple potential syntactic encodings for equivalent semantic sentences, what factors govern the use of one form rather than another. While the Uniform Information Density hypothesis has been popular in computational psycholinguistics literature over the last decade, I argue that the empirical evidence supporting it needs to be re-evaluated. A general framework of Incremental Generation makes convergent predictions in many cases and a more powerful set of explanations overall. Owing to the modular nature of the language production system, any factors which correlate with lexical access time or the speed of constituent construction will also be correlates of output linear order. This includes conditional ‘Information Density’ as well as a range of other properties including frequency, definiteness, and length. In this paper, I provide background on the above issues and propose the verb-particle alternation as an informative window into theories of language production. Statistical modeling over an extremely large sample of natural language data shows that, to whatever degree we can characterize the output of the language production system as ‘efficient’ in information ordering, this is an emergent property of an incremental generation system.

Keywords: Language production; Incremental generation; Syntactic optionality; Verb-Particle construction; Statistical modeling.

*Correspondence concerning this article should be addressed to Spencer Caplan. E-mail: spcaplan@sas.upenn.edu
1 Language Production

1.1 Incrementality

That the planning and realization of speech occurs at a number of hierarchical levels is a widely supported view with a long history (Fromkin [1971], Lashley [1951], Wundt [1904]). More recently this is popularly divided into a few major processes: conceptualization, grammatical encoding (also called ‘formulation’ on some accounts (A. Roelofs, 1997; A. P. A. Roelofs, 1992)), and phonological encoding (or articulation) (Bock & Levelt, 2002).

Conceptualization can be thought of as generating and mapping some intention onto an abstract ‘message’. This results in some conceptual information which still needs to be encoded linguistically in order to accomplish the speaker’s communicative goal. The message contains the required lexical concepts and the relationship between them.

The heart of the language production system is in grammatical encoding (V. Ferreira & Dell, 2000). Grammatical encoding starts with a message (or part of a message which is being continually developed) and outputs the word forms which make up the phonological content handled by the articulatory-phonetic system. Inside of grammatical encoding, a process retrieves the linguistic units which correspond to these lexical concepts and assigns them required thematic and functional structure (Levelt, 1992, 1993). Such lemmas are then linearized in accordance with syntactic restrictions. The end result is an articulatory plan for the utterance which can be realized as overt speech. This general framework is diagrammed in Figure 1.

What is important here is not only the presence of such distinct sub-systems to language generation, but the sequentially ordered relation between them (F. Ferreira & Swets, 2002). The most crucial organization is that language production functions incrementally (F. Ferreira & Swets, 2002; Pechmann, 1989; A. Roelofs, 1997, 2013; Smedt, 1990). Only after a piece of information is made available at the immediately higher level of processing does it become available to trigger activity at the next level down the production stack. This means that a word cannot begin being assigned phonological structure until it has received a semantic role. Phonetic articulation of an element cannot begin until it has been assigned morphological form, etc. While it may seem intuitive, this ‘vertical’ incrementality might be contrasted with a view in which a syntactic tree is first built, with individual words being slotted in afterwards (see V. Ferreira (1996) for support of incrementality of lexical retrieval over ‘competition’ models). Additionally, the system as a whole functions in parallel. Once a lemma has been sent off for morphological inflection, the system does not need to wait in order to start retrieving subsequent lemmas. “At the same time that a piece of information works its way from idea to articulation, other pieces are constructed and make their way through the system as well.” (F. Ferreira & Swets, 2002). Patterns in speech errors strongly indicate that functional and positional processing are able to function in parallel (Dell, 1985, Hoppe-Graff, Herrmann, Winterhoff-Spurk, & Mangold, 1985). While various levels of production can operate in parallel, individual elements within the system (e.g. lemmas) are constrained sequentially and incrementally.

Such assumptions about the incremental nature of language production have wide empirical support with evidence ranging from natural speech errors, to elicited productions in paired-association and picture naming tasks, etc. (Brown-Schmidt & Konopka, 2015; F. Ferreira & Swets, 2002).

1.2 Timing

Because language production is incremental, variations in the order in which information is delivered from one component to the next affect the order in which element appear in speech. Since the output of higher-level processing systems feeds the input of lower-level system, this view of incremental production naturally predicts that higher-level modules do not need to complete their work on an utterance before the next level begins. We might term this behavior ‘Minimized Buffering’ because without it, whatever lemmas are retrieved first would need to wait in a buffer for slower elements before continuing down the stack. There are explicit computational accounts of this principle (Smedt, 1990) as well as empirical support from visual-world naming tasks (Pechmann, 1989). As described in Smedt (1990):

“Natural speech is often produced in a piecemeal fashion: speakers start to articulate a sentence before the syntactic structure, or even the meaning content of that sentence
has been fully determined. Under the assumption that the human language processing apparatus is capable of carrying out different tasks in parallel, the speaker may already utter the first words while simultaneously processing more content to be incorporated in the sentence. If the next fragment is ready to be uttered when the first fragment nears completion, speech is produced fluently and without hesitation. This mode of generation, which is called incremental generation, seems to serve a system whose major purpose is to articulate relatively fluent speech, even if it is imperfect or incomplete. Once a partial sentence has been constructed, the generator will try to complete the sentence in a maximally grammatical way. There is less load on short term memory if concepts can be uttered roughly as they become accessible.”

An Incremental Generation (IG) framework (adopted by many, but explicitly implemented in e.g. Kempen and Hoenkamp (1987); Smedt (1990)) builds pieces of phrase structure as the lemmas become available. Minimized Buffering means that fragments are fit together as quickly as possible so long as constraints of phrase structure are not violated. Again, the final structure is built in a ‘piecemeal and heuristic fashion under the control of lemmas and their functions, rather than by means of an algorithm that generates a tree into which words must be inserted.’ (Bock & Levelt 2002).

When this IG framework is taken together with the notion of Minimized Buffering, certain constituent ordering predictions are made. Specifically, whatever factors correlate with initial conceptual retrieval or faster lexical access should correlate with that constituent being linearized first in the resultant sentence. An elucidating example of this is given in V. Ferreira and Dell (2000):

“Assume a speaker wishes to describe the outcome of the race between the tortoise and the hare in Aesop’s fable with a verb such as defeat. Furthermore, assume that the lemma for the word hare is quickly activated and selected. Given the early selection of the hare lemma, the most efficient strategy is for the speaker to produce the passive, The hare was defeated by the tortoise, rather than the active, The tortoise defeated the hare, since only with the passive can the already-selected hare lemma be immediately mentioned. If the speaker produces the active, then one of two inefficient processing strategies must be adopted: Either the already-selected hare lemma must remain active in a buffered state until the sentence-final position arrives for production (while other words are selected and produced in earlier sentence positions), or the already-selected hare lemma must be deactivated and subsequently reactivated. From this, a general principle can be induced: Production proceeds more efficiently if syntactic structures are used that permit quickly selected lemmas to be mentioned as soon as possible.”

This type of mechanism is additionally supported by work on the link between visual/conceptual processing and linear order (Bunger, Papafragou, & Trueswell 2013; Gleitman, January, Nappa & Trueswell 2007), which manipulated not the lexical access speed for words, but the order in which concepts were activated through the visual system and found a similar effect on output production order.
A summary of the Incremental Generation framework for this article is given below.

<table>
<thead>
<tr>
<th>Incremental Generation Account of Optionality</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Module responsible for lexical retrieval is incremental</td>
</tr>
<tr>
<td>• Operation between different modules (functional vs. positional assignment) functions in parallel</td>
</tr>
<tr>
<td>• Minimized buffering (Lemmas are not held in buffer longer than required by constraints of syntax)</td>
</tr>
<tr>
<td>• Any factors which speed up lexical access are also proxies for output linear order.</td>
</tr>
</tbody>
</table>

2 Optionality

While speakers generally produce utterances in a spontaneous, fluent, and seemingly effortless fashion, a rich cognitive architecture underlies a complex chain to move from concepts to spoken sentences. A difficulty in attempting to study the mechanisms of language production is that, most of the time, the largest contributor to spoken output is the meanings which speakers want to convey in particular contexts. As such, any theory would attempt to make the same predictions in the majority of cases, i.e. output the grammatical and attested sentences while being unable to output the ungrammatical or otherwise impossible sentences. Following Garrett (1975), much work in language production has focused on speech error phenomena.

However, in wanting to ensure that theoretical predictions made on the basis of processing failure generalizes to ‘typical’ cases as well, another major testing ground for accounts of language production is the study of ‘syntactic optionality’: given multiple potential syntactic encodings for equivalent semantic sentences, what factors govern the use of one form rather than another. Additionally, speech errors might be caused at a number of different modules within the language production stack. Since the tasks that induce speech errors often add general processing difficulties, it is tough to isolate the role of grammatical encoding in particular rather than phonological encoding through speech errors alone.

Simply declaring that “the field should investigate syntactic optionality in production” is trickier than we would like however, as many apparent cases of ‘optionality’ may actually carry wildly different pragmatic implicatures. Consider the difference in pragmatic interpretation between (1a-1b).

(1) a. I made a mistake.
    b. Mistakes were made.

Thus instances like the active/passive alternation may not be ideal cases of ‘optionality’. In natural settings, context and communicative intent play a major role in the use of one form over the other. This is especially true if the speaker carries meta-awareness of the varying pragmatic differences as in [1b]—which is likely an intentional hedge against the idea that the speaker is the one who has made mistakes.

Probably the most well-studied case of syntactic alternation is ‘that’-omission (V. Ferreira & Dell, 2000; Jaeger, 2010; Lasnik & Fiengo, 1974), a type of syntactic alternation like shown in (2a-2b).

(2) a. The coach said the players were tired.
b. The coach said that the players were tired.

In certain sentences containing complement clauses (some complications to be discussed in Section 3), the grammar allows apparent optionality with respect to whether or not the complementizer ‘that’ is overtly pronounced or not. Various production and discourse accounts have been put forth to explain what drives the optional omission/realization of ‘that’ in such cases, as discussed below.

2.1 Ambiguity Avoidance

In order for linguistic communication to be successful, a hierarchical syntactic representation needs to be both encoded by the speaker as a linear sequence of words as well as successfully decoded by the listener back into a hierarchical representation. While a given sequence of words may correspond to a multitude of structurally sound representations, most of the time contextual information limits interpretation to only a single plausible interpretation. Nonetheless, real-time language processing functions not over an entire sentence in batch, but over an incrementally received sequence of partial information over time. Such ‘garden paths’ (Traxler & Pickering, 1996; Trueswell, Tanenhaus, & Kello, 1993) occur when an utterance contains a temporary syntactic ambiguity that is biased toward a syntactic analysis that will eventually turn out to be wrong. Particularly when extra-syntactic properties (context, frequency, etc.) lead to a strong garden path, it temporarily hinders communication (F. Ferreira & Henderson, 1991; Trueswell, Sekerina, Hill, & Logrip, 1999).

It is possible that the language production system involves some computational oversight in order to behave in such a way that limits garden paths and inefficient output (Clark & Tree, 2002; Hankamer, 1973; Temperley, 2003). Consider the following alternation as it relates to ‘that’-omission:

(3) a. The coach knew you well and thought you’d be a good candidate.
   b. The coach knew you missed practice regularly

Even though the first four words are the same between 3a and 3b, the eventual syntactic interpretation of ‘you’ is not. In 3a it needs to be interpreted as a direct object, while in 3b it is an embedded subject. After only hearing ‘The coach knew you...’ however, it is ambiguous which interpretation to assign. This arises in this case because this particular pronominal is phonologically identical in both subject and object interpretations. Other pronouns such as ‘I, she, he’, etc. occur in complementary distribution with respect to subject versus object roles (e.g., I vs. me). If the language production system were sensitive to avoiding temporary ambiguities, then we would expect the use of ‘that’ to be more frequent when the beginning of the complement clause contains an ambiguity (like with ‘you’) compared to when it does not (like with ‘I/he/she’).

Experiments in V. Ferreira and Dell (2000) tested precisely this claim and found no evidence of implicit ambiguity avoidance through syntactic optionality. This family of experiments is discussed further in Section 2.2.

2.2 Sentence-Recall

V. Ferreira and Dell (2000) tested predictions of an ambiguity avoidance account of optionality (Temperley, 2003) against an Incremental Generation account (as described in Section 1). The
particular methodology used was a specific type of targeted elicitation known as a “sentence-recall” task (Potter & Lombardi, 1990). In the sentence-recall paradigm, participants are first given a list of sentences (typically on the order of three). Participants are then asked to recall each sentence in production based on a limited number of cue words taken from that sentence. For example, if the target sentence were ‘The coach knew you well and thought you’d be a good candidate.’, this might be cued with ‘coach’ and ‘candidate’. The motivation for sentence-recall is that while recall for general semantics is very good, memory for specific syntactic encoding is quite a bit poorer than one might expect. Subjects are only marginally above chance at mirroring the target syntax of a cued sentence.

In the particular implementation by V. Ferreira and Dell (2000), target sentences contained embedded subjects that were either morphologically ambiguous (e.g. ‘I know you’) or subjected to repetition priming (Forster & Davis, 1984) (e.g. ‘I know I’). Since prior mention of a lemma is expected to speed re-activation, it is predicted by Incremental Generation to correlate with more ‘that’-omission.

The general findings of V. Ferreira and Dell (2000) offer support for an Incremental Generation framework and no evidence of ambiguity avoidance. However, the sentence recall task is not a typical conversational setting, and compromises naturalness in pursuit of experimental control. V. Ferreira and Hudson (2011) use a variant task involving natural dialogue and find no effect of prior mention.

Two possibilities are likely contributing here. First, sentence-recall is fundamentally a memory task, and does not engage the same components of sentence generation that typical communication does. Additionally, it may be the case that the effect of prior mention on lexical access times is being overshadowed by host of other factors. Both points raise an important question of generalizability, and motivate the study of natural speech (which is not subject to methodological confound in the same way) and at a large scale (which allows for the simultaneous comparison of multiple competing factors and hypotheses). This is the approach taken by this paper, as well as Jaeger (2010) which we discuss below.

### 2.3 Uniform Information Density

A prominent previous account, the ‘Uniform Information Density’ hypothesis (UID) (Jaeger, 2010; Levy & Jaeger, 2007), proposes that syntactic optionality is driven by a speaker’s implicit managing of computable information content to maximize communicative efficiency. This view starts from the notion that if language has a primary function of communication, we might imagine that each utterance conveys some particular ‘amount of information’. The hypothesis of Uniform Information Density (UID) is that, agnostic to the actual implementation of the language processing system, it must have an upper bound to the amount of information it can process within a fixed time. In order for communication to be efficient, the speaker would not want to convey too much information at once (which would be difficult to process) nor would he/she want to produce speech that is overly redundant (since that is a waste of potential bandwidth for information transfer). As languages always allow some degree of optionality of expression for given semantico-pragmatic contents, then in order for language processing to be efficient the amount of information conveyed over time should be relatively more uniform than non-uniform. In this way there is an intuitive relationship between ‘Information’ and ‘predictability’: The more
an event can be reasonably expected, the less we have learned upon its occurrence. Conversely, if an event is assumed very unlikely, then its occurring the listener a great deal of new information. Jaeger (2010) takes an inverse log transformation over conditional probability to serve as a representative proxy for information (Eq. [1]). Note that this is not the intuitive notion of ‘information’ as ‘some meaningful propositional content’ but rather a particular estimate of Shannon information (Shannon, 1948). All that means here is that ‘Information’ is simply the inverse log transformation of predictability/probability of some word conditioned on a previous word.

\[ \text{Information(word)} = \log \frac{1}{p(\text{word})} = -\log p(\text{word}) \]  (1)

This brings us to a prose definition of UID: “Within the bounds defined by grammar, speakers prefer utterances that distribute information uniformly across the signal (information density). Where speakers have a choice between several variants to encode their message, they prefer the variant with more uniform information density (ceteris paribus)” (Jaeger, 2010)

The application to ‘that’-omission follows from the fact that individual predicates vary in terms of how likely they are to introduce a complement clause. A verb like ‘thinks’ is very likely to introduce a complement clause (CC), and so hearing the overt complementizer ‘that’ provides listeners with relatively little new information. Conversely ‘confirmed’ is notably less likely to occur with complement clauses, and so hearing ‘we’ directly follow it is a fairly surprising event (since the morphology indicates it must be an embedded subject). A schematic of this idea is shown in Figure 2.

Jaeger (2010) tests the UID hypothesis by extracting a sample of complement clause instances (approx. 7,000) —1,173 (17.5%) of which have a complementizer, while 5,543 (82.5%) do not—from the Penn Treebank (Marcus, Marcinkiewicz, & Santorini, 1993) subset of the Switchboard corpus of telephone dialogues (Godfrey, Holliman, & McDaniel, 1992). A multiple regression model is used to predict the binary outcome of overt ‘that’. The regression included a number of independent factors intended as proxies for different theories of language production, including some which can be given a coherent interpretation under IG. Some of these are discussed in turn below.

Since the CC directly follows the matrix verb in the majority of sentences, then processing difficulties at the matrix verb may manifest at the beginning of the CC. As it becomes more difficult to process the matrix verb, then speakers should be more likely to overtly produce the complementizer (as it is already active by the time the verb was retrieved from memory) (Roland, Elman, & Ferreira, 2006). As will be discussed further in Section 5.1, less frequent words are slower to retrieve from memory (Bock & Levelt, 2002).

The same relationship might then hold for the frequency of the CC-subject; likelihood of ‘that’-omission would go up as the frequency of the CC-subject goes down, since those are slower to retrieve from memory.

As Information Density goes up, that means that predictability in context goes down, which correlates with ‘that’-mentioning. The interpretation of this factor under IG, rather than UID, is discussed in Sections 3 and 5.2.

The intuition behind ‘prior mention’ follows from previous sentence-recall experiments (V. Ferreira & Dell, 2000). Lemmas that were recently retrieve from memory are faster to activate again. The faster it is to build the CC-subject the less likely it should be to overtly include ‘that’.

As the CC-subject constituent grows longer, it is slower to construct before phonological output. This should favor ‘that’-mentioning since it’s increasingly likely that the complementizer will have
been activated before the entire CC-subject is constructed.

Speech rate factors like a pause directly before the CC, or general disfluency is a proxy for general processing difficulty in accessing the CC-subject or other upcoming material. This should positively correlate with ‘that’-mentioning.

The resultant analysis is that, while many factors correlate with ‘that’-omission, Information Density is a significant predictor as well, even after controlling for alternate theories. A non-exhaustive summary of the main results in Jaeger (2010) is shown in Table I.

This finding to support UID (and the ‘efficient language’ framework more broadly) has generated a large audience and resultant body of work with claims ranging from language production (A. F. Frank & Jaeger, 2008), and speech rate (Pellegrino, Coupé, & Marsico, 2011), to word order typology (Maurits, Navarro, & Perfors, 2010), distributional properties of the lexicon (Plantadosi, Tily, & Gibson, 2011), and language acquisition (Fedzechkina, Jaeger, & Newport, 2012). With such apparent centrality to language, it should be important to understand a production
Table 1: Sample summary output from Jaeger (2010)’s analysis of ‘that’-omission. Dependent variable is the presence of overt ‘that’. Most, although not all, production-level factors correlate with ‘that’-output.

| Factor                          | Estimate | Std. Error | Z-value | P(>|z|) |
|--------------------------------|----------|------------|---------|---------|
| Frequency (Matrix Verb)        | -0.23    | 0.03       | -7.7    | ~0      |
| Frequency (CC-Subject)         | -0.02    | 0.03       | -0.7    | >0.5    |
| Information (CC | Matrix Verb) | 0.47      | 0.03     | 16.9    | ~0      |
| CC-subject prior mention       | -0.32    | 0.17       | -1.9    | <0.052  |
| CC-subject length               | 0.18     | 0.014      | 12.8    | ~0      |
| Pause                           | 1.11     | 0.11       | 10.2    | ~0      |
| Disfluency                      | 0.39     | 0.12       | 3.2     | <0.002  |

In the next section, I argue that despite previous modeling of ‘that’-omissions, findings on UID need to be re-evaluated.

3 Re-evaluating Support for UID

The original finding from Jaeger (2010) however requires re-evaluation. While a log transformation over conditional probability is taken to be the proxy for information content on Jaeger’s approach, the fact that such conditional probability (and contextual predictability more generally) should correlate with the output of grammatical optionality is not a unique prediction of UID. On Jaeger’s model, the major factor encoded to represent ease of lexical access is frequency. But this is more akin to a strawman than a legitimate interpretation of previous language production theories, e.g. V. Ferreira and Dell (2000); Smedt (1991).

Despite the claim that “Given the definition of information, UID assumes that speakers have access to probability distributions over linguistic units (segments, words, syntactic structures, etc.). This distinguishes UID from most existing production accounts, which make different architectural assumptions and do not predict information density to affect speakers’ preferences’ (Jaeger 2010), there is no reason why an IG account in incompatible with probability sensitive representations. In fact, predictability is a strong predictor of lexical access times (Ehrlich & Rayner 1981; Rayner 1998; Rayner, Ashby, Pollatsek, & Reichle, 2004; Staub 2011) as much as raw word frequency is (Inhoff & Rayner 1986; Rayner & Duffy 1986). This effect of predictability (also commonly called ‘surprisal’ in the parsing literature) is robustly attested across a number of different methodologies[1]. Scanpaths during reading are more irregular when predictability is low (von der Malsburg, Kliegl, & Vasishth 2015). Within studies of event related potentials, predictability is negatively correlated to the amplitude of the N400 component (S. L. Frank, Otten, Galli, & Vigliocco 2015). The timecourse of activation in functional MRI is correlated with predictability values estimated from language models using both surface n-grams (Willems, Frank, Nijhof, Hagoort, & Van den Bosch 2015) and more rich phrase-structure (Henderson, Choi, Lowder, & Ferreira 2016). Given this, an Incremental Generation framework also predicts that,

---

[1] See Hale (2014) for a review of such effects and effective models in sentence comprehension
since conditional probability (predictability) is tied to access times, it should also be a correlate of ‘that’-mentioning.

It is worth investigating the idea that generally local computation by an incremental language production mechanism is responsible for the sort of trends that have alternatively been put forth in favor UID or communicative optimization schemes. UID is not a causal mechanism, but rather a high level description about optionality. As such it requires a disjunctive explanation of why factors like frequency effect system output and why predictability effects the same output. I will argue that we achieve a better understanding of the system through IG, which offers a unified mechanism underlying both the UID effects as well the general connection between lexical access times and conceptual access ordering.

An additional issue is to be raised regarding the generalizability of ‘that’-omission in studying language production. The ability of researchers to make generalizations about internal behavior of the processing mechanism relies on the assumption that choices are not being governed by important (but potentially difficult to detect) differences in social factors or syntactic representation. For instance, while the order of agent and patient constituents in the active/passive alternation is flipped, it is clear that the use of one form over another is strongly restricted by a semantic/pragmatic distinction. While some alternations are apparently innocuous — ‘The coach said the players are lazy’ vs. ‘The coach said that the players are lazy’ — the lexical semantics of different matrix predicates not only biases toward one syntactic surface realization or another, but in many sub-cases the ‘choice’ is not in fact an option at all (Grimshaw, 2009).

Rates of ‘that’-omission vary a great deal by register. Elness (1984) finds that in scientific writing only 1.3% of complement clauses had no that. Whereas in “adventure fiction and westerns” that same rate was 58.1%. In natural conversation, the percentage of sentences with that-deletion rises to about 85% (Biber, 1999; Elness, 1984). The data which Jaeger (2010) extracted from Switchboard exhibits a ‘that’-omission rate of 82.5%, generally in line with the high-rate reported for conversational speech. However, we need to proceed with caution when rates of omission in Switchboard are roughly 60 times higher than in the formal register. With such a high degree of variability in ‘that’-omission attributable to sociolinguistic register, it is not clear how much we can expect to learn about the cognitive generation mechanism.

Taken together, these points motivate considering an alternative ‘optional’ construction over which theories of language generation might be tested. Section 4 discusses the verb-particle alternation as such a case and tests the predictions of UID and Incremental Generation in this application.

### 4 Verb-Particle Construction

I propose to investigate a different syntactic alternation from ‘that’-omission, the verb-particle construction. This is beneficial because there are potentially fewer lexi-co-semantic confounds compared with the predicate classes of matrix verbs generally. It shows far less variability by register. And there are more degrees of freedom to evaluate model predictions (relations taken between both object and particle with respect to both verb, and each other). By extracting a large, diverse sample of verb-particle alternation data from naturally occurring speech and text, we are able to test predictions of various theories of language production on this novel domain.

The syntax of such verb-particle constructions is well-researched (Aarts, 1989; Farrell, 2005), although the particular details don’t weigh heavily on the present work. Consider the alternation
This alternation includes a transitive verb, a morphologically invariant word which we might call a ‘particle’, and a direct object noun phrase. The sentence in 4a shows what we will term ‘particle-first’ order, while 4b exhibits ‘object-first’ order. The verb-particle alternation in (4) should not be conflated with the superficially similar construction such as in (5a). While the basic word order between (4) and (5a) appears similar, the underlying structure is quite dissimilar, as we can see by lack of the word order alternation as in (5b). The particles in verb-particle constructions are not prepositions (i.e. they do not take objects), though words used as particles tend to be used as prepositions as well. Also, note that pronominal objects are unacceptable in particle-first order as in 6.

It is worth acknowledging work addressing the semantic details of various verbal particles as in (Brinton, 1985; Dehé, 2002; Gries, 2003). It is an oversimplification to say that such verb-particle sentences are all in the same semantic category as they might be divided into (at least) the following buckets (see Dehé (2002) for more information): Compositional as in 7, Idiomatic as in 8, and Aspectual as in 9.

However, for now I will proceed without attempting to further classify verb-particle semantics in this case.

4.1 Data Extraction

Verb-particle alternation instances were extracted from the Corpus of Contemporary American English (COCA) (Davies, 2009). This was an an ideal source as it is extremely large, allowing us to test predictions on very specific interactions between theoretical factors. Additionally, it consists of five distinct genres (subcorpora): Academic texts, speech, news, magazines, and news. We
see that, unlike in the case of that-deletion, only a relatively minor fraction of total variance is explained by register differences across genre (see\(^2\)). The overall rate of particle-first order ranges only between 65% and 85%. This is far smaller than the amount of sociolinguistically conditioned variance present in the case of ‘that’-omission, which ranges between 1% and 85% Biber (1999); Elness (1984).

Figure 3: Bar plot of particle-first order by genre in COCA. While some variability by register is present, this is far more limited than the sociolinguistically conditioned variance present in the case of ‘that’-omission

While some previous work has attempted to automatically extract instances of verb-particle constructions (Baldwin, 2005; Kim & Baldwin, 2006), such research has been intended for use in downstream NLP applications rather than attempting to answer theoretical cognitive questions. Candidate verbs were identified using pronoun construction (because we see as possible ‘John picked him up’, but not *’John picked up him’). Thus, if a verb could take part in a verb-particle construction, then we would expect to see it attested taking a pronoun object given a sufficiently large sample. This is easy to search for without getting false positives. We search for all phrases matching the form ‘\textsc{verb pronoun particle’}. We subsequently identified instances of such verbs containing neither pronouns nor prepositional phrases, e.g. ‘John put down the book down, but not *’John put down on the table the book’. Output data was tagged for a number of relevant factors including: order of particle, frequencies, conditional probabilities, constituent length, definiteness, source corpus, etc.\(^3\) These statistical measures were estimated over the entirety of COCA rather than only the verb-particle sentences which our regression model will eventually attempt to predict.

Data was limited to verbs which appeared as taking an object/particle combination (in either order) at least ten times and were not exclusively categorical in their particle-first or particle-second order.\(^4\) Data were additionally filtered by running all extracted sentences through the Stanford lexicalized PCFG parser (Manning et al., 2014). We discarded any cases for which the parsed output did not contain exactly a phrasal verb followed by two daughter constituents tagged as a noun phrase (NP) and particle (PRT) in either order. After cleaning up there are 67,905

\(^2\)The code for performing this data extraction and subsequent statistical analysis is available open-source at https://github.com/scaplan/PhrasalVerbs-ProcessingAccount

\(^3\)This helps remove idiomatic expressions which are not underlyingly optional in output order
unique sentences to be predicted. The size of the extracted data is valuable since it allows for the evaluation of multiple hypotheses even over extremely rare events. These include 54,955 particle-first and 12,950 particle-second sentences. The total ratio of particle-first order is approx. 81%. This general ratio of particle-first is a replication of previous work on manual extraction of verb-particle cases (Kroch & Small, 1978). There are 296 unique verb-particle pairs. This corresponds to 99 unique verbs over 33,085 unique triples (tuples of verb, particle, nounHead).

4.2 Data Quality

Any automated syntactic annotation or extraction scheme will necessarily contain some number of errors. No gold-standard corpus of verb-particle sentences exists against which the present extraction scheme can be evaluated. To evaluate the approximate quality of extracted data I took a sample of 100 tagged instances and manually judged their status (true potential verb-particle alternation rather than a prepositional phrase for instance (probably want to refer to above examples or something). On this sample, approximately 90% of cases were true positive examples of the verb-particle alternation. The 10% of error cases are sentences which do not allow the alternate constituent order for syntactic reasons rather than facts of the processing mechanism. These were fairly evenly split between particle-first and object-first order. For example see below:

(10) a. We pushed our way out.
    b. *We pushed out our way
    c. We pushed out a quality research project

(11) a. Why do I have to get out a period later than you?
    b. *Why do I have to get a period later than you out?
    c. Let’s get the birthday cake out.

Such cases are difficult to automatically avoid since the verb-particle pairs themselves can surface in the optional order some of the time. But in these particular instances are only available on an alternate semantic interpretation which does not allow re-ordering.

The syntactic annotation scheme used for the Penn tree-bank was not designed with such fine-grained lexical semantic distinctions in mind. It’s a necessary consequence that, based on these parse trees alone, some limited errors in construction identification arise. This would be true whether extraction were performed based on hand-annotated structures or algorithm-based. So long as the rate of errors is relatively low and unbiased, we can simply acknowledge it and proceed without it interfering in subsequent analysis.

5 Explanatory Modeling of IG and UID

When writing statistical models like those in this paper, it is crucial to uphold the idea that a model which predicts linguistic output with high accuracy is not, in and of itself, an explanatory
theory. It is insufficient to model the outcome of syntactic optionality (or any phenomenon) unless we can move towards understanding why correlated variables have predictive power. Statistical tools can only verify, but not produce, empirical hypotheses. (Yang, 2008)

In this section, I review the major independent variables included in the subsequent regression. For each correlation, we should think back to the potential mechanisms that could have generated such output (and particularly when accounts share predictions or diverge).

When a speaker is producing a sentence that will contain a verb-particle construction, the moment immediately following the output of the verb is an important inflection point. The speaker still needs to output both the relevant object and the relevant particle, but at this stage, the syntax allows those two elements to be linearized in either order. This triggers what is essentially a ‘race’ between the two constituents. If the object is retrieved and constructed first then it will be linearized first, and vice versa for the particle. Any factors which speed up lexical access will also be proxies for output order.

(\textit{NOTE: I should draw a clear flowchart to illustrate the above ‘race’ at the post-verbal inflection point})

Incremental Generation offers an explanations of why we see what might alternatively be termed ‘Uniform Information Density’. UID results not from a central governing principle for communicative efficiency, but as an emergent property from the algorithmic implementation of language production in the mind. Lemmas are retrieved as triggered by the intention to express various concepts. Not every lemmas is retrieved at with the same speed. Since the system is incremental, but partially parallel (between modules) then variations in lemma access are manifested as comparable variations in linear order when the grammar permits.

5.1 Frequency

Higher frequency words have faster average retrieval times from memory (Bock & Levelt, 2002; Forster & Chambers, 1973; Paap, McDonald, Schvaneveldt, & Noel, 1987). Under an IG account, if the frequency of the object and the particle are pitted against one another: As frequency of an element goes up, so should the likelihood that it is linearized first.

The following plot shows the mean inverse-log frequency of the object in particle-first order sentences compared with particle-second sentences. Since the object constituent commonly consists of multiple words, the frequency computation is made over the head noun. The inverse log frequency can be thought of as the additional ‘processing cost’ resulting from lower frequencies; so as frequencies go down then the inverse-log (cost) goes up. Less frequent objects correlate with particle-first order, and vice versa for particles (Figure 4). This effect is confirmed via t-test for both objects ($t = 30.021$, $df = 21594$, p-value < 2.2e-16) and particles ($t = -54.615$, $df = 15937$, p-value < 2.2e-16).

This effect of frequency is a unique prediction of IG that is not made by UID. Even though UID is a theory defined over Shannon information rather than frequencies, it is simple enough to think through what would happen if we reconceptualized UID as ‘uniform frequency density’. Since frequencies are stable and unconditioned values, then the gap between particle frequency and object frequency is constant regardless of the order with which the words are linearized. The uniformity of the flow of lexical frequencies over time does not depend on linear order and so no uniform ordering theory (whether defined over Shannon Information or frequencies) makes

\footnote{For example, in a longer constituent like ‘the big green ball’ this means that the operative measure is the total number of occurrences of ‘ball’}
explicit predictions with respect to frequency and linear order. This is not to say that UID is incompatible with attested frequency effects, simply that UID is orthogonal to them; it neither predicts nor offers an explanation of frequency.

Under IG, higher frequency correlates with earlier linear order because those elements are, on average, faster to retrieve from memory—a high frequency object is more likely to win the linearization race. More specific mechanisms for this connection, in particular Serial Rank Access (Lignos, 2013; Murray & Forster, 2004), are discussed further in Section 6.3.

5.2 Predictability

Rather than computing the baseline likelihood of a word or constituent overall (raw frequency, as in Section 5.1) we might estimate the conditional frequency or predictability of a word in a particular local context. This captures important information because, for instance, the probability of ‘the’ is high overall (it is the most frequency word in English), yet the likelihood that ‘the’ occurs directly following ‘The coach that the...’ is almost zero. An inverse-log transformation over predictability is precisely the statistical proxy for Information content used under UID (Eq. 2)

\[
\text{Information}(\text{object}|\text{verb}) = \log \frac{1}{p(\text{object}|\text{verb})} = -\log p(\text{object}|\text{verb})
\]

This paper takes two measures of predictability at the inflection point after the verb has been produced: P(object|verb) and P(particle|verb). Conditioning on the verb tabulates total occurrences of either the object or the particle within at least five words to the right of the verb. As with the case of unconditioned frequency, and following Jaeger (2010), the ‘object’ is actually a tabulation of the noun head rather than a multi-word phrase.
Computing the average $\text{Information}(\text{object} \mid \text{verb})$ and $\text{Information}(\text{particle} \mid \text{verb})$ in particle-first vs. particle-second sentences, we see that higher object information (lower predictability and thus higher processing cost) does in fact correlate with particle-first order and vice versa for particles (Figure 5). This effect is confirmed via a t-test for both objects ($t = 22.694$, $df = 20462$, p-value $<$ 2.2e-16) and particles ($t = -59.53$, $df = 21538$, p-value $<$ 2.2e-16).

Figure 5: Mean inverse-log predictability (information density) for elements in particle-first and particle-second sentences. Grey bars indicate standard error

Under IG, these measures of predictability are predicted to correlate with linear order for the same reason that frequency does. As the predictability of the object in context is lower (and hence information density higher), then lexical access times are correspondingly slower. The slower it is to retrieve and construct the object, the more likely it is for the particle to win the linearization race and be sent off for positional processing first (and vice versa). As noted in Section 3, predictability is a strong correlate of lexical access times (Ehrlich & Rayner, 1981; Rayner, 1998; Rayner et al., 2004; Staub, 2011). This effect is measurable in reading speed (Rayner, 1998), distribution of scanpaths (von der Malsburg et al., 2015), as well as several neurolinguistic measures (S. L. Frank et al., 2015; Henderson et al., 2016; Willems et al., 2015). While the connection between predictability and processing speed is a robust empirical effect, IG is not tied to any particular mechanism which enables this connection to emerge.

UID makes a convergent prediction here on the relationship between conditional predictability and linear order, but based on a different set of assumptions. Unlike in the case of ‘that’-omission, where the complementizer is understood by Jaeger (2010) to convey no additional information aside from a syntactic relation, the particle in these verb-particle constructions actually may convey a great deal of semantic information depending on the predicate. In fact, UID requires that predictability of an element increases monotonically as a function of the prior output string of words. The intuition being that whatever initial asymmetry in predictability there is between $P(\text{particle} \mid \text{verb})$ compared with $P(\text{object} \mid \text{verb})$, then whichever is linearized first increases the
baseline predictability of the second element. By mentioning the more predictable element first, that would reduce the asymmetry in information content at each time, while mentioning the low predictability element first would increase the asymmetry in information. For example, imagine that after producing the word ‘pick’ then the predictability of ‘up’ is 0.5 and the predictability of ‘book’ is 0.1. If the next word uttered is in fact ‘up’ that the subsequent predictability of ‘book’ would rise higher, say to 0.6; while the verb imposes some selectional restrictions on the object, it would seem natural that a verb-particle pair would further reduce the upcoming search space making predictability greater. Even if the selectional restrictions placed on the particle by the object are on average less drastic, it should be intuitive that predictability doesn’t get reduced based on additional information. So while the gap in predictability directly following the verb was 0.4 (0.5 for ‘up’ minus 0.1 for ‘book’), the eventual gap in predictability would be reduced if the more predictable element were linearized first (‘up’ is output at 0.5 predictability while the predictability of ‘book’ rises to 0.2, hence the gap is reduced to 0.3) and it would be increased if the less predictable element were linearized first (‘book’ is output at 0.1 predictability while the predictability of ‘up’ rises to 0.6, hence the gap is increased to 0.5).

If that assumption of monotonically increasing predictability were dropped, then UID would make no prediction regarding speakers’ ordering of these elements in the verb particle construction, as is the case with unconditioned frequencies. On the other hand, an IG account requires no such assumption of monotonicity in order to predict a correlation between higher predictability and linear order; the relation is mediated through lexical access times.

5.3 Downstream Effects and Predictability

Rather than conditioning predictability solely on the verb, \( P(\text{object}|\text{verb}) \) and \( P(\text{particle}|\text{verb}) \), we can also compute the downstream effects of production after the inflection point. For instance, if the object were linearized first then \( P(\text{particle}|\text{object}) \) would immediately become relevant, or if the particle were linearized first then \( P(\text{object} | \text{particle}) \) would be of note. These ‘downstream’ predictability values may be hugely asymmetric and hence have an effect on total constituent construction.

Predictions under both IG and UID are convergent here — both accounts predict that as \( P(\text{particle}|\text{object}) \) goes up, that should correlate with object-first order and vice versa — albeit for the different reasons explicated in 5.2.

5.4 Definiteness

Definite articles presuppose identifiability or familiarity within a context or discourse. This should correlate with faster lexical access times, and thus an IG framework predicts that definite objects should be more likely to occur in object-first constructions compared with indefinites. Constructions containing definite article occur with particle-first order 82.8% of the time, compared with 77% for indefinite objects.

5It is worth noting that the case of pronominal objects is, in spirit, similar to the effect of definiteness under an IG account. Since pronouns are highly frequent and, in context, refer to some typically salient individual their lexical access speed would in general be quite fast, consistent with object-first order. However, the apparent categoricity of the pronoun output order results from this being grammaticalized, which leaves underspecified the mechanism by which this grammaticalized happened historically and how pronoun order is stably acquired now.
This is unsurprising given previous studies of definiteness and linear order (Ransom, 1977). Another view is that definiteness serves as a reasonable proxy for the given vs. new status of conceptual information within a discourse. In the English dative ordering alternation, Bresnan, Cueni, Nikitina, Baayen, et al. (2007); Collins (1995) find an overwhelming effect of discourse status (given vs. new) on constituent ordering.

While it is not possible for us to estimate the discourse status of constituents directly from a corpus with unknown speakers and contexts, definiteness is a reasonable, albeit limited, proxy for such discourse structure.

5.5 Object Length

Noun phrases consisting of more words contain more pieces, and should take more time to build within the production system. IG thus predicts object length to correlate with particle-first order based on the slowdown in time to construct the constituent. The effect of phrase length on constituent ordering is well-studied (Kimball, 1973; Stallings et al., 1998). Since information content, following Jaeger (2010), is computed between the head of the noun phrase object rather than other elements, UID makes no explicit prediction regarding the role of object length on linear order.

The median length (in words) of objects within particle-first sentences is five, compared to a median length of three words in object-first instances. The significance of the difference between these object-length distributions is confirmed by a t-test (t = 36.35, df = 23512, p-value < 2.2e-16).

5.6 Prior Mention

Repeated mention of a constituent has been shown to facilitate faster lexical retrieval (Forster & Davis, 1984). Previous work on ‘that’-omission has seen mixed effects on this regard; V. Ferreira and Dell (2000) found a strong effect within a sentence-recall paradigm, while that effect disappeared under more naturalistic dialogue settings (V. Ferreira & Hudson, 2011). A positive marginal effect are present in Jaeger (2010) (see Table 1).

Here we measure ‘prior mention’ as a categorical outcome when the noun head of the verb-particle object had occurred previously in the same sentence.

If repeat mention does significantly speed up lexical access, then an IG framework predicts that it should correlate with object-first order, while this is orthogonal to UID. We find a notable correlation as the proportion of particle-first sentences is approximately 3% lower under object repeat-mention cases compared to the base condition (0.772 compared to 0.807). Likewise particle-first order is more likely in particle repeat-mention cases compared to the base condition (0.839 vs. 0.807).

---

6The alternation illustrated in (1a) vs. (1b).

(1) a. Santa gave [toys] [to the children]
   b. Santa gave [the children] [toys]

7In natural dialogue it would be more natural to repeat only the head rather than an entire object verbatim. Contrast the naturalness of (1a) compared to (1b).

(1) a. Find the biggest and most brightly colored apple in the display, and now hand me the apple.
   b. Find the biggest and most brightly colored apple in the display, and now hand me the biggest and most brightly colored apple.
compared to 0.807). One possibility for this difference between the previous data on ‘that’-omission and the present verb-particle data is that repeat-mention is a very rare occurrence. Only about 1.7% of verb-particle sentences include object repeat-mention, while that figure is only 0.5% for particle repeat-mention. Previous marginal effects might result not because there isn’t an underlying effect, but simply because this paper represents the first instance in which a sufficient magnitude of data was analyzed to detect the effect.

**6 Results**

A multilevel logit model, a type of generalized linear mixed model (Breslow & Clayton, 1993), was used to evaluate the factors representing predictions of the IG framework of language production compared with the UID hypothesis. The dependent variable was the analyzed binary outcome of linear order (particle-first ordering rather than object-first). Predictions of IG and UID are summarized in Table 2.

<table>
<thead>
<tr>
<th>Variable</th>
<th>IG Prediction</th>
<th>UID Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency(object)</td>
<td>Positive</td>
<td>Agnostic</td>
</tr>
<tr>
<td>Frequency(particle)</td>
<td>Negative</td>
<td>Agnostic</td>
</tr>
<tr>
<td>Information(object</td>
<td>verb)</td>
<td>Positive</td>
</tr>
<tr>
<td>Information(particle</td>
<td>verb)</td>
<td>Negative</td>
</tr>
<tr>
<td>Information(object</td>
<td>particle)</td>
<td>Negative</td>
</tr>
<tr>
<td>Information(particle</td>
<td>object)</td>
<td>Positive</td>
</tr>
<tr>
<td>Object prior mention</td>
<td>Negative</td>
<td>Agnostic</td>
</tr>
<tr>
<td>Particle prior mention</td>
<td>Positive</td>
<td>Agnostic</td>
</tr>
<tr>
<td>Object Length</td>
<td>Positive</td>
<td>Agnostic</td>
</tr>
<tr>
<td>Definite Object</td>
<td>Negative</td>
<td>Agnostic</td>
</tr>
</tbody>
</table>

Table 2: Predictions from IG and UID on verb-particle output order

The model includes the aforementioned independent predictors as ‘fixed-effects’, as well as a random effect variable for each individual verb as well as each genre (subcorpus). Results of the model are shown in Table 3.

It is important to know what we aim to learn from examining the output of such a logit model. The goal here is not an engineering evaluation, and so we are not concerned directly with minimizing the variance left unexplained by the model. In fact, increasing the number of variables would generally tend to improve model fit —encoding random effects for each verb-particle pair as well as every unique object would certainly improve model fit. But such a model would not teach us about which aspect of the underlying sentence production system are
responsible for the attested variance. The goal here is not to write the ‘best’ model (meaning the one that strictly captures the most variance). What we care about is understanding why different components capture variance to begin with. The object of study is the individual fixed effects rather than explicit model optimization. It is additionally important that in the simple case without any regularization, the variable independent factors exist on different scales with different distributions. Thus it is not straightforward to use the magnitude of the coefficients as a direct proxy for effect size. We are primarily concerned with the direction of effects (sign of the estimates) and corresponding significance.

| Factor                      | Estimate  | Std. Error | Z-value | P(>|z|) |
|-----------------------------|-----------|------------|---------|---------|
| (Intercept)                 | 9.728563  | 0.301967   | 32.217  | ~0      |
| Frequency(object)           | 0.321612  | 0.012811   | 25.105  | ~0      |
| Frequency(particle)         | -1.281312 | 0.026418   | -48.501 | ~0      |
| Information(object | verb)     | 0.099888   | 0.008528 | 11.713  | ~0      |
| Information(particle | verb)     | -0.156983  | 0.019035  | -8.247  | ~0      |
| Information(object | particle) | -0.340724  | 0.012058  | -28.256 | ~0      |
| Information(particle | object)   | 0.207724   | 0.009239  | 22.482  | ~0      |
| Object prior mention        | -0.321580 | 0.090248   | -3.563  | 0.000366|
| Particle prior mention      | -0.067802 | 0.177409   | -0.382  | 0.702330|
| Object Length               | 0.908835  | 0.020169   | 45.062  | ~0      |
| Definite Object             | -0.730315 | 0.026545   | -27.512 | ~0      |

Table 3: Output of primary logistic regression model. Dependent variable is particle-first order. Random effects of individual verb and genre omitted for brevity. ‘~0’ indicates a value less than 0.00001

With the exception of ‘particle prior mention’, we see a strong significant correlation between every independent variable predicted by IG and the dependent linear order. 

6.1 Controls and Robustness

When dealing with uncontrolled observational data, it is imperative to properly control for noise and underlying bias. Including additional random effects improves overall model fit, but the fixed
effects remain stable even when a random slope is introduced into the model for each verb-particle pair (296 pairs in total). See Table 4.

| Factor                                | Estimate  | Std. Error | Z-value | P(|z|) |
|---------------------------------------|-----------|------------|---------|-------|
| (Intercept)                           | 12.490327 | 2.177339   | 5.737   | ∼0    |
| Frequency(object)                     | 0.370373  | 0.014501   | 25.542  | ∼0    |
| Frequency(particle)                   | -1.534307 | 0.236814   | -6.479  | ∼0    |
| Information(object | verb)     | 0.117150   | 0.009377  | 12.494 | ∼0    |
| Information(particle | verb)     | -0.340232  | 0.096603  | -3.522 | 0.000428 |
| Information(object | particle) | -0.384688  | 0.013568  | -28.352 | ∼0    |
| Information(particle | object)  | 0.177314   | 0.009041  | 19.612 | ∼0    |
| Object prior mention                  | -0.347215 | 0.096451   | -3.600  | 0.000318 |
| Particle prior mention                | -0.017521 | 0.186990   | -0.094  | 0.925346 |
| Object Length                         | 0.976255  | 0.022189   | 43.996  | ∼0    |
| Definite Object                       | -0.832507 | 0.029289   | -28.423 | ∼0    |

Table 4: Output of logistic regression including additional random effects (296 individual verb-particle pairs as well as an effect for each genre). Dependent variable is particle-first order. ‘0’ indicates a value less than 0.00001. While the model fit is greatly improved by including such values as random effects (BIC of 39559.1 with additional random effects compared to 44932.8 without) the direction and significance of fixed effects remains consistent.

The results were additionally tested for robustness by performing a k-fold cross-validation experiment ([Picard & Cook, 1984]) with largely consistent results (See Appendix A for full results and further discussion).

One concern in dealing with complex regression involving numerous independent factors is collinearity. If predictors (or sets of predictors) are correlated with one another, then it is especially difficult to interpret regression output or establish clear theories of causality.

6.2 Object Length Experiment

The argument from [Jaeger, 2010] in support of UID is that, even after controlling for correlates of other processing theories, the effect of UID manifests in the regression model. Additionally, Jaeger attempts to make a direct comparison of the relative theoretical status between such factors by examining the size of coefficients for such actors:

“To put the effect in relation to two theories of sentence production that have received considerable attention in psycholinguistic work on syntactic variation, availability-based sentence production and dependency processing accounts: the effect associated with the only parameter fitted for information density outranks the effect of all three parameters associated with dependency length effects in the model. The effect of information density also is much larger than the combined effect of accessibility related parameters in the model. That is, information density emerges as the single most important predictor of complementizer that-mentioning.”
However, comparing the coefficient sizes between variables is not necessarily a good way to judge their relative ‘status’ in the processor. Because more sentences happen to contain short objects (Figure 6, it is somewhat necessarily the case that the effect of object length appears smaller. This alone does not serve as evidence that we should afford UID elevated status within our conception of a production model.

By looking at the subset of sentences with somewhat longer objects we give a fairer chance to compare the two theories. On UID the conditional predictability really is central as a proxy for information density. If speaker’s manage syntactic information density as posited, then the effects of predictability on linear order should remain present in medium length objects. This is not to say we shouldn’t expect the coefficient size to change, just that UID predicts information density to remain a significant factor in the regression model.

Alternatively on the IG account, predictability is only one of many factors which correlate with linear order by way of having an effect on lexical access times. On this view, there is no particularly special status for conditional probability compared to frequency compared to object length. Each factor has some effect on lexical retrieval speed (potentially in opposing directions). On this equal footing, we should expect the predictive role of information density to disappear when looking at longer objects —any boost in retrieval speed that increased predictability may give would be overshadowed by the increased amount of time it takes to build and process a multi-word object.

To compare these views, I re-ran similar regression analyses as above but limited to sentences whose objects are at least N words long for various values of N. Evaluating cases of N=2 or more words there are 58,628 instances (at a ratio of 82.05% particle-first).
Table 5: Evaluating cases of N=2 or more words. This is qualitatively the same as evaluating all cases.

However, the picture changes when we look at even medium length objects —N=4 or more words (5,688 instances at a ratio of 96.6% particle first)

Table 6: Evaluating cases of N=4 or more words. The effect of conditional probability is absent, while the effects of frequency, object length, and definiteness remain.

When restricted to evaluating even medium length objects (four or more words), then there is no effect of information density. Importantly, while the proportion of particle-first cases is approaching categorical here (driven by the length of the object), other processing factors remain significant predictors of linear order. Note that definiteness and frequency are still significant factors. This is strong direct evidence against the predictions of UID, and consistent with an IG framework.

If speakers really privileged (implicitly) managing information density, then they would need to buffer retrieved lemmas more than they appear to do given the current data and model. To be clear, UID doesn’t require that there be no effect of object length. it just simply doesn’t make a prediction on its own that there should be an effect either way. Not does it offer an explanation of such attested correlations. To say UID has explanatory power over the correlation with predictability is to require an additional theory to account for the remaining effects of frequency, length, etc. Whereas on an IG account, the discussed correlations are all accounted for and predicted in a unified way. UID is a descriptive tendency of an aspect of the dataset and is not a cause of any of it. Optionality should be a source of evidence, not a phenomenon to be explained in its own right. Information theory is not a silver bullet. The information theoretic effects that pop up in the system importantly interact with other factors, and may disappear under the right (or perhaps wrong) circumstances.
6.3 Rank, Mutual Information, and Interactions

It would be ideal to be able to ground predictive factors in specific mechanisms of action. For instance, while the propositions ‘frequency correlates with access times’ and ‘frequency correlates with linear order’ are true empirically, it is natural to wonder why and how this might be the case. If frequency were a direct contributor to linear order, then there could be an arbitrary shape to the relationship between frequency and the verb-particle output; the frequency relation could be linear or logarithmic or any number of possible mappings. However, on the IG account frequency is only a correlate of verb-particle output insofar as it serves as a proxy for lexical access speed. Since it is well-established that the relationship between frequency and lexical access is logarithmic—captured under a Rank-based serial access mechanism (Lignos, 2013; Murray & Forster, 2004)—IG specifically predicts that log frequency should be a better correlate of verb-particle output than raw frequency. This prediction is tested via a model comparison analysis in Table 7.

One potential concern for any information theoretic account (including UID) is that some variables are rather direct measures—the number of words in a constituent doesn’t require any parameters for a modeler to set—whereas other variables such as predictability are(noisy) estimates of the abstract concepts they are intended to represent. While conditional probability as calculated from surface n-grams empirically correlates well with a number of performance measures (lexical access time (Staub, 2011), scanpath variance (von der Malsburg et al., 2015), neurolinguistic measures (S. L. Frank et al., 2015; Willems et al., 2015), etc.), it is possible that the variability of correlation between information density and verb-particle output is limited by implementing an estimate which correlates with, but is crucially different than, what the production system actually monitors.

One alternative information theoretic implementation would be to compute pointwise mutual information ($\text{pmi}$) (Fano, 1961). That is, rather than simply looking at the predictability of some word or structure in a given context we might examine the relative change in predictability. Intuitively this means that rather than asking ‘How likely am I to encounter the object ‘book’ now that I’ve heard ‘pick’?’ we ask ‘How much more likely am I to encounter the object ‘book’ now that I’ve hard ‘pick’ compared to my general representation of the frequency of book?’ For applications of $\text{pmi}$ in natural language processing and psycholinguistics see e.g. Bouma (2009), Church and Hanks (1990), Jaeger and Weatherholtz (2016). The prediction on this alternative account is that the boost in predictability should grow as the raw frequency of the term goes down. While Jaeger (2010)’s implementation of Information Density is agnostic to frequency, and the Incremental Generation account predicts both higher frequency and higher predictability to work in the same direction (correlating with first-mention), the $\text{pmi}$ account predicts that predictability has an inverse relationship with frequency.

$$\text{pmi}(\text{obj}; \text{verb}) = \log \frac{p(\text{obj}, \text{verb})}{p(\text{obj})p(\text{verb})} = \log \frac{p(\text{obj}|\text{verb})}{p(\text{obj})}$$

To evaluate both the frequency/Rank prediction as well as this alternative implementation of an information theoretic account, we performed model comparisons between the effect of the individual measures of frequency, predictability (information), and $\text{pmi}$ (which ties together the relationship between frequency and predictability). We ran several comparable multilevel logit model as before; each model attempts to predict the binary outcome of linear order (particle-first ordering rather than particle-second). Each model includes a single fixed effect representing either
frequency, object predictability, or object-verb PMI as well as object prior match, object length, definiteness, and random effects for genre and verb. See Table 2. This type of pairwise model comparison is informative, but we need to keep in mind the caveat of colinearity. Even in a model which does not include a term for predictability, because frequency is colinear, we can really only withhold the ‘residual information’. Directly interpreting resultant model behavior should proceed with caution.

<table>
<thead>
<tr>
<th>Model Delta</th>
<th>BIC</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>60923</td>
<td>Log Frequency (Rank)</td>
</tr>
<tr>
<td>145</td>
<td>61068</td>
<td>Raw Frequency</td>
</tr>
<tr>
<td>348</td>
<td>61271</td>
<td>Information(object</td>
</tr>
<tr>
<td>434</td>
<td>61357</td>
<td>PMI</td>
</tr>
<tr>
<td>604</td>
<td>61527</td>
<td>Raw object predictability</td>
</tr>
</tbody>
</table>

Table 7: Output of model comparison between PMI, predictability and frequency. Bayesian Information Criterion (BIC) [Schwarz et al., 1978] is a standard measure of model quality. Smaller BICs indicate better models. The model includes the fixed effects of the listed variable, object prior match, object length, definiteness, and random effects for genre and verb

Investigating the relative model quality of regressions fit using only frequency, predictability, or PMI we see no benefit of PMI as an alternative information theoretic implementation over simple information density. This is unsurprising given that PMI is symmetric. It doesn’t apply well to the case of language production which is an inherently dynamic, and linearly ordered string. Simply because ‘the math works’ does not mean that it generalizes well to a particular linguistic problem. For instance, one might wonder what is interaction term is between frequency and predictability. But such a question isn’t well formed with respect to the verb-particle construction. The verb will always be uttered before either the object or the particle, and so we can’t ask what the output of the system would be when the predictability term is removed. But even without knowing every fine grained detail of the internal workings of the language production system, we can still make inferences about the families of models that could be operating.

The model fit using log frequency outperforms the other models, including raw frequency or predictability alone. This is generally consistent with a Rank-based serial access mechanism [Murray & Forster, 2004]; log frequency is a better proxy for rank than raw frequency. Additionally, frequency on its own outperforms predictability, which is consistent with IG framework of language production. Predictability has an effect on lexical access times, and thus on constituent ordering, but that is evidently, at least partially, overtaken by the effect of frequency. Log predictability, like log frequency, accounts for more variance than raw predictability, but there is not a clear interpretation of this result [8]. Given the effects of both frequency and predictability, it is thus descriptively accurate to say that the general output of language production is more uniform with respect to information ordering at syntactic choice points than it might otherwise be by chance. However, this is an emergent property of IG. The effects of predictability on production interact with a host of other factors such a frequency, definiteness etc. Not only is there no reason to privileged the role of predictability over other factors in production.

---

8It is potentially worth considering that taking the log transform of a heavily skewed distribution like frequency or predictability results in a distribution that is much closer to Gaussian, but there is nothing inherent to logistic regression that would bias against non-normally distributed independent variables.
7 Conclusion

A large body of work in psycholinguistics is concerned with the mechanisms underlying the language production system (Kempen & Hoenkamp 1987; Levelt 1993; Smedt 1990), i.e. what is the cognitive architecture managing the link from concepts to spoken sentences. A major testing ground for accounts of language production is the study of ‘syntactic optionality’; given multiple potential syntactic encodings for equivalent semantic sentences, what factors govern the use of one form rather than another (V. Ferreira & Dell 2000).

While the ‘Uniform Information Density’ hypothesis (UID) (Jaeger 2010), has attracted a good deal of attention in the computational psycholinguistic literature over the last decade, the empirical evidence supporting it deserved re-evaluation. Notably, a realistic interpretation of existing models of incremental language production (V. Ferreira & Dell 2000; Smedt 1990) make convergent predictions in many cases. This is true because predictability (the same value used as a proxy for information density) is a known correlate of lexical access time (Staub 2011).

As evidence for theories of optionality has, to date, largely come from ‘that’-omission is it important to note the complexities of the selectional restrictions of many embedding predicates (Grimshaw 2009). Additionally, rates of ‘that’-omission show such high variability by register (Biber 1999; Elsness 1984) that it is not clear what we might learn about the cognitive architecture of the production system.

By instead modeling data from the verb-particle construction, we are able to more fully evaluate the Uniform Information Density hypothesis with respect to Incremental Generation. Statistical modeling by multilevel logit models shows strong support for a range of lexical access predictors including the effect of conditional probability posited by both UID and Incremental Generation. However, when limited to evaluation over even moderately long objects (at least four words), then the predictions of UID are not borne out.

There is not direct evidence here for any mechanism behind the hypothesis of Uniform Information Density. At best this is more akin to a statistical description of output data rather than explanation for any underlying generation mechanism. To whatever degree we can characterize the output of the language production system as ‘efficient’ in information ordering, this is an emergent property of a simple, incremental generation system.

Acknowledgments

I would like to thank Tony Kroch for many important discussions throughout this project. Also thank you to Charles Yang, John Trueswell, Mitch Marcus, Ryan Budnick, and João Sedoc.

References


Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. D. (2004). The effects of frequency and


Wundt, W. (1904). The psychology of the sentence. *Language and Psychology: Historical Aspects of
A  Cross Validation

Rather than evaluating a regression over the entire data set, we randomly divided the data into ten equally sized subsets and subsequently fit a the same regression to each of them (Table 8). The cross-validation experiment is useful in providing a measure of stability; since any subset of the data introduces the possibility of a biased sample, we can estimate which factors are robust in light of the reduced size and increased noise. However, this type of cross-validation results should be considered with the possible caveat of under-sampling. Since the object repeat-mention variable only applies to a small portion of the data to begin with, it is possible that the effect loses significance on an individual sample not because it is underlyingly invalid, but simply because there were too few relevant instances to detect the effect. Likewise if a sampled subset of the data happened to include a high proportion of long object cases, then we expect the Information density effect to drop out (as per Section 6.2).

A safe interpretation of the cross-validation experiment is that it increases our confidence in the attested robust effects, but it does not necessarily invalidate effects which occasionally disappear (like repeat mention or predictability)

| Variable                  | Proportion Significant Runs | Mean Pr(>|z|) |
|---------------------------|----------------------------|--------------|
| Frequency(object)         | 1.0                        | ∼0           |
| Frequency(particle)       | 1.0                        | ∼0           |
| Information(object | verb)         | 0.9                        | 0.02202303    |
| Information(particle | verb)         | 1.0                        | 0.000569499   |
| Information(object | particle)      | 1.0                        | ∼0           |
| Information(particle | object)      | 1.0                        | ∼0           |
| Object prior mention      | 0.2                        | 0.202379      |
| Particle prior mention    | 0.0                        | 0.5282941     |
| Object Length             | 1.0                        | ∼0           |
| Definite Object           | 1.0                        | ∼0           |

Table 8: K-fold cross validation output (k=10)

B  Colinearity

Assesing Baseline: Variance Inflation Factor (VIF) is a standard measure of diagnosing the potential problems induced by colinearity, but VIF cannot be straightforwardly calculated for logistic regression. It would be useful to write out a full table in the appendix measuring the $R^2$ for each variable predicted by the others.

Ridge Regression: Regardless of the degree of colinearity, the question is if it is affecting the reported results. A sanity check for this is comparing the base output to a Ridge regression. But I haven’t done that yet since it requires properly rescaling the variables (and that’s not straightforward when some are gradient and others are categorical). In the best case scenario there is not a notable different in my basic logistic regresison output and the ridge regression output, but that’s an empirical question.
Problem identification: Lastly, another approach is to plot the outcome of each individual explanatory variable and compare to ground truth for the sentences I’ve manually labeled. Are the outliers the same as the manually tagged error cases? If so, then there’s little concern. If not, then I need to go through and blacklist those problematic verb-particle pairs from the dataset.