Investigating Locality Effects and Surprisal in Written English Syntactic Choice Phenomena

Rajakrishnan Rajkumar[a]
Marten van Schijndel[b]
Michael White[c]
William Schuler[d]

[a]Department of Humanities and Social Sciences, IIT Delhi, Hauz Khas, New Delhi, India 110016, raja@iitd.ac.in (corresponding author)
[b]Department of Linguistics, The Ohio State University, Oxley Hall, 1712 Neil Ave., Columbus, OH 43210 USA, vanschm@ling.osu.edu
[c]Department of Linguistics, The Ohio State University, Oxley Hall, 1712 Neil Ave., Columbus, OH 43210 USA, mwhite@ling.osu.edu
[d]Department of Linguistics, The Ohio State University, Oxley Hall, 1712 Neil Ave., Columbus, OH 43210 USA, schuler@ling.osu.edu

Abstract

We investigate the extent to which syntactic choice in written English is influenced by processing considerations as predicted by Gibson's (2000) Dependency Locality Theory (DLT) and Surprisal Theory (Hale 2001, Levy 2008). A long line of previous work attests that languages display a tendency for shorter dependencies, and in a previous corpus study, Temperley (2007) provided evidence that this tendency exerts a strong influence on constituent ordering choices. However, Temperley's study included no frequency-based controls, and subsequent work on sentence comprehension with broad-coverage eye-tracking corpora found weak or negative effects of DLT-based measures when frequency effects were statistically controlled for (Demberg & Keller 2008, van Schijndel, Nguyen, & Schuler 2013, van Schijndel & Schuler 2013), calling into question the actual impact of dependency locality on syntactic choice phenomena. Going beyond Temperley's work, we show that DLT integration costs are indeed a significant predictor of syntactic choice in written English even in the presence of competing frequency-based and cognitively motivated control factors, including n-gram probability and PCFG surprisal as well as embedding depth (Wu, Bachrach, Cardenas, & Schuler 2010, Yngve 1960). Our study also shows that the predictions of dependency length and surprisal are only moderately correlated, a finding which mirrors Demberg & Keller's (2008) results for sentence comprehension. Further, we demonstrate that the efficacy of dependency length in predicting the corpus choice increases with increasing head-dependent distances. At the same time, we find that the tendency towards dependency locality is not always observed, and with pre-verbal adjuncts in particular, non-locality cases are found more often than not. In contrast, surprisal is effective in these cases, and the embedding depth measures further increase prediction accuracy. We discuss the implications of our findings for theories of language comprehension and production, and conclude with a discussion of questions our work raises for future research.

Index terms— language production, dependency locality, surprisal, constituent ordering

1 Introduction

A long line of previous research, comprising both spontaneous production experiments and corpus analyses, has studied the production biases involved with constituent ordering. In general, languages are attested to favor producing shorter dependencies, as Liu (2008) demonstrates in a cross-linguistic study involving twenty languages. Figure 1 shows this trend for English using data from two corpora, the Brown corpus (Francis & Kučera 1989) and the Wall Street Journal (WSJ) portion of the Penn Treebank (PTB; Marcus, Marcinkiewicz, and Santorini 1993).

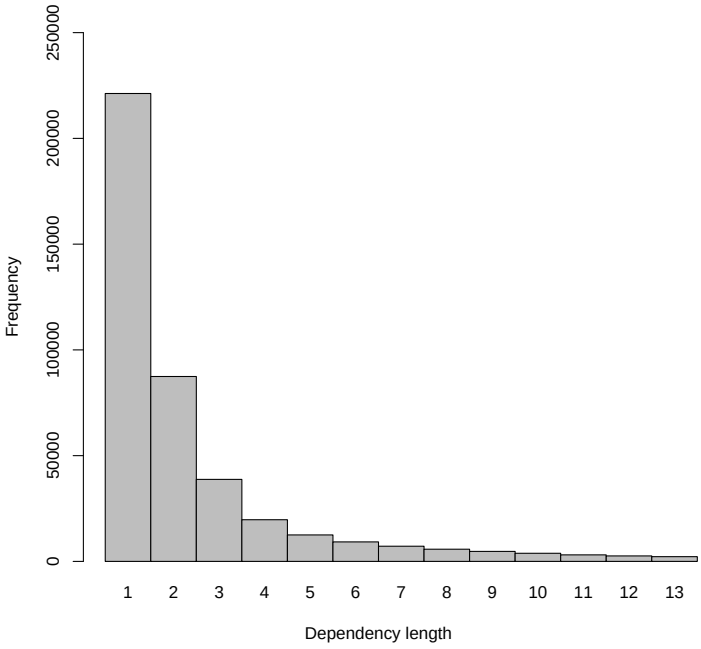
In this paper, we investigate whether this generalization holds true for constructions where speakers have a choice of expressing the same idea using competing word orders, as in the following example (italics added):

(1) a. One day Maeterlinck, coming *with a friend* upon an event which he recognized as the exact pattern of a previous dream, detailed the ensuing occurrences in advance so accurately that his companion was completely mystified. (Brown corpus CF03.10.0)

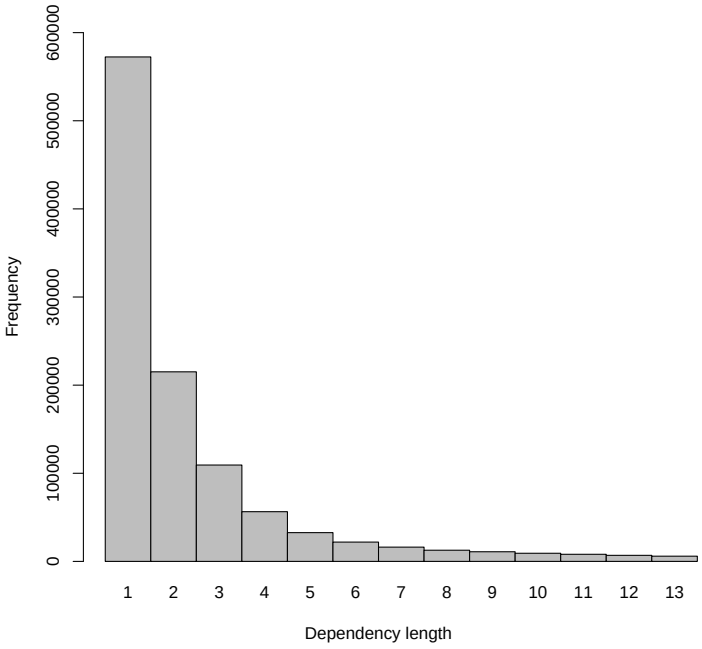
b. One day Maeterlinck, coming upon an event which he recognized as the exact pattern of a previous dream *with a friend*, detailed the ensuing occurrences in advance so accurately that his companion was completely mystified. (Constructed alternative)

Research in the past decade has investigated the hypothesis that one of the factors which influences the structuring of languages is the ease of comprehension and production, in addition to abstract learning biases in language acquisition (Chater & Christiansen 2010, Hawkins 2004; 2014). More concretely, do speakers display a preference to produce (1-a) above, since it is easier to produce or comprehend compared to (1-b)? Using a corpus study, Temperley (2007) showed that the tendency to minimize DEPENDENCY LENGTH has a strong influence on constituent ordering choices in written English. In the corpus sentence (1-a), there is a short intervening adjunct *with a friend* between the verb *coming* and the subsequent long constituent starting with *upon*, thus inducing a shorter dependency in comparison to the competing order in the constructed alternative (1-b). Moreover, it is easy to misparse the variant as having *previous dream with a friend* as a constituent, even though this gives rise to a nonsensical interpretation where the dream is a joint activity with the friend.

Dependency length minimization has a long history in the literature dating back to Behaghel's (1932) principle of end weight. In a long line of



(a) Brown corpus



(b) WSJ corpus

Figure 1: Dependency length distributions

pioneering work, Hawkins has shown that languages tend to prefer shorter dependencies (Hawkins 1994; 2000; 2001; 2004; 2014). In the context of syntactic choice phenomena like heavy NP shift (Arnold, Wasow, Losongco, & Ginstrom 2000, Wasow 2002), dative alternation (Bresnan, Cueni, Nikitina, & Baayen 2007), verb-particle shifts (Hawkins 2011, Lohse, Hawkins, & Wasow 2004) and topicalization and left-dislocation (Snider & Zaenen 2006), many other works also corroborate the tendency of languages to minimize dependency length. There is cross-lingual evidence that word order patterns in SOV languages conform to dependency locality (Hawkins 1994; 2004). The definition of Early Immediate Constituents (EIC) in Hawkins (1994) predicts that for verb-final languages, long constituents tend to precede short ones in the preverbal position. He validates his prediction using Japanese data, and subsequent research builds on EIC predictions in language production studies in Japanese (Yamashita & Chang 2001) and Korean (Choi 2007). There is also parallel evidence from optional function words, which are likely to be omitted to shorten dependencies (Hawkins 2001; 2003, Jaeger 2006; 2010; 2011).

Temperley's (2007) corpus study uses a variant of Gibson's Dependency Locality Theory (DLT; Gibson 1998; 2000), a resource-limitation theory of human sentence comprehension, to account for a wide variety of syntactic choice constructions in two written English corpora. Crucially, Temperley's corpus study does not control for other possible explanations of syntactic choice aside from DLT; in particular, it includes no frequency-based controls. Explaining syntactic choice data in terms of a single factor (viz. length or dependency minimization) has also been criticized as being reductive (Bresnan et al. 2007, Snider & Zaenen 2006, Wasow 2002). Some corpus studies on specific constructions either hold frequency constant or control for it with lexical counts, as in the case of studies on Heavy NP shift (Arnold, Wasow, Asudeh, & Alrenga 2004, Arnold et al. 2000), dative alternation (Bresnan et al. 2007), object relative clauses (Jaeger 2006), complement clauses (Jaeger 2010) and subject relative clauses (Jaeger 2011). These studies provide preliminary evidence that dependency length is a significant predictor of ordering choices even when frequency-based controls are considered.

However, in sentence comprehension, although dependency length has been shown to correlate with reading times on constructed stimuli (Levy, Fedorenko, & Gibson 2013, Warren & Gibson 2002), it has been difficult to replicate this effect in broad-coverage naturalistic data as strong statistical frequency controls reduce or reverse the effect of dependency length (Demberg & Keller 2008, Shain, van Schijndel, Gibson, & Schuler 2016, van

Schijndel et al. 2013, van Schijndel & Schuler 2013).[1] Even when previous production studies have used explicit frequency controls, they have only used frequency information about individual lexical items and the frames those items occur in, which may not be sufficient. For example, van Schijndel, Schuler, and Culicover (2014) demonstrated that the structural bias statistics captured by latent-variable PCFGs are at least as strong a frequency confound in comprehension as the information captured by lexical counts and subcategorization frame frequencies. Importantly, the structural biases they examine stem from underlying syntactic configurations which may not be readily apparent when counting the number of times a given lexical item occurs in a certain frame (e.g., the probability of a gap being passed into a left branch compared with a right branch at each point in the syntax tree is independent of any lexical item and would require an impractically large norming study to manually control for). Since structural statistics may also confound studies of locality's influence on sentence production, this work uses stronger frequency controls than previous production studies by statistically controlling for both structural and lexical information.

This paper extends Temperley's work by testing the hypothesis that **dependency length is a significant predictor of syntactic choice in written English even in the presence of competing frequency-based and cognitively grounded control factors**. Recent work in computational psycholinguistics has used information-theoretic measures to model both language comprehension as well as production. From the perspective of language comprehension,[2] one of the factors hypothesized to represent comprehension difficulty is SURPRISAL (Hale 2001, Levy 2008), which quantifies the predictability of a word in a given linguistic context. More predictable words induce faster processing times in reading (Boston, Hale, Patil, Kliegl, & Vasishth 2008, Demberg & Keller 2008, Smith & Levy 2013). Thus surprisal as a control variable models the extent to which the text is comprehensible. In addition, we use EMBEDDING DEPTH (Wu et al. 2010, Yngve 1960) as a control, since increased memory depth is considered to increase comprehension difficulty.

[1] As one of the reviewers pointed out, frequency can be considered as an interesting factor in its own right. Please refer to Table 6.1 of MacDonald (1999) which points to many works which consider frequency in production and comprehension research.

[2] We choose controls in part from the sentence comprehension literature since the editing done by careful authors may take comprehensibility considerations into account explicitly (Jaeger 2011). Additionally, in early Natural Language Generation (NLG) work, editing done by the author is considered equivalent to self-monitoring in Levelt's (1989) model of human language production (Neumann & van Noord 1992).

To date, corpus studies of constituent ordering choices have developed separate analyses for each construction investigated. For example, in the model of dative alternation presented by Bresnan et al. (2007), the logistic regression model (Breslow & Clayton 1993) predicts the choice of obtaining *NP-NP* vs. *NP-PP* objects for each verb. In a methodological advance upon this study and other previous corpus studies cited above, we develop analyses involving a variety of constructions in the same model. To do so, following the technique described by Joachims (2002) for reducing ranking to pairwise classification, we train the logistic regression model to predict the corpus choice over other constructed grammatical variants, rather than predicting whether the corpus choice is of a particular form (e.g. *NP-NP*). The technique of training a ranking model to prefer the corpus variant over other alternatives is common in the natural language generation (NLG) literature (Rajkumar & White 2014). Indeed, using this technique, White and Rajkumar (2012) have shown that including total dependency length in an otherwise comprehensive ranking model yields significantly improved ordering choices in NLG. Their work provides preliminary evidence for the efficacy of dependency length as a predictor of syntactic choice amidst other competing structural and lexical factors, though using a more complex setup than employed here, which does not permit the statistical significance of predictors to be easily assessed.

In this paper, we show that for constituent ordering across a variety of constructions in written English, the minimal dependency length theory of language comprehension (Gibson 2000) is indeed a significant predictor of the corpus choice even in the presence of competing frequency-based and cognitively grounded controls (n-gram log probability, latent variable PCFG surprisal and embedding depth measures) proposed in the computational psycholinguistics literature (Demberg & Keller 2008, Roark, Bachrach, Cardenas, & Pallier 2009, Wu et al. 2010), in particular for various postverbal syntactic choice alternations. We also investigated the extent to which the aforementioned controls accounted for cases which diverged from the dominant tendency of English to observe locality constraints (non-locality cases). Surprisal and dependency length are only moderately correlated and their predictions model disparate parts of the data, with surprisal correctly predicting many non-locality cases. We report that embedding depth measures collectively induce significant increases in the prediction accuracy of non-locality cases over a frequency-based baseline involving n-gram probability and PCFG surprisal.

As Arnold (2011) discusses in detail, sentence production theories account for production phenomena either via constraints or processes inherent

to the production system (speaker-internal as in Arnold et al. 2000, Ferreira 2003) or resorting to explanations where constraints on comprehension influence language production (listener-oriented as in Branigan, Pickering, & Cleland 2000, Clark & Haviland 1977). However, it is difficult to separate speaker and listener-oriented processes in language production. Hawkins' efficiency principles are also compatible with both speaker and listener-oriented perspectives (Hawkins 2011). Since our study is based on written data, we avoid committing to either of these explanations, as writers and editors are actively engaged in maximizing the comprehensibility of the text for the benefit of the readers. We leave open the possibility of future studies involving spoken data to make a definitive statement. Moreover, as Jaeger and Buz (in press) discuss, speaker-internal and listener-oriented explanations need not be mutually exclusive. Reflecting this observation, they adopt the labels *production ease* (MacDonald 2013) and *communicative accounts*, where the latter label avoids the implication that communicative aspects are solely for the benefit of the listener. Consistent with this view, we find it plausible that speakers may over time learn to make choices in particular circumstances that lead to effective communication without having to engage in costly real-time reasoning about the competing possibilities.

The rest of the paper is structured as follows. Section 2 provides the requisite background for the study and Section 3 discusses the relationship between dependency length and other factors influencing constituent ordering. Section 4 describes our data and Section 5 presents the results of our experiments. Subsequently, Section 6 provides a discussion of Dependency Locality in the context of our results. Section 7 reflects on the implications of our findings for theories of language comprehension and production. Finally, Section 8 summarizes the conclusions of the study and discusses questions our work raises for future research.

2 Background

This section provides detailed background on Dependency Locality Theory (DLT; Gibson 1998; 2000) and Surprisal Theory (Hale 2001, Levy 2008), two influential theories of sentence processing that we use in this work. DLT was originally proposed as a theory of resource limitation explaining the complexity of unambiguous structures (subject and object relative clauses). This study also investigates the extent to which non-DLT measures of processing complexity can predict syntactic choice. While DLT predicts an influence from the length of dependencies, increased memory load may also reduce

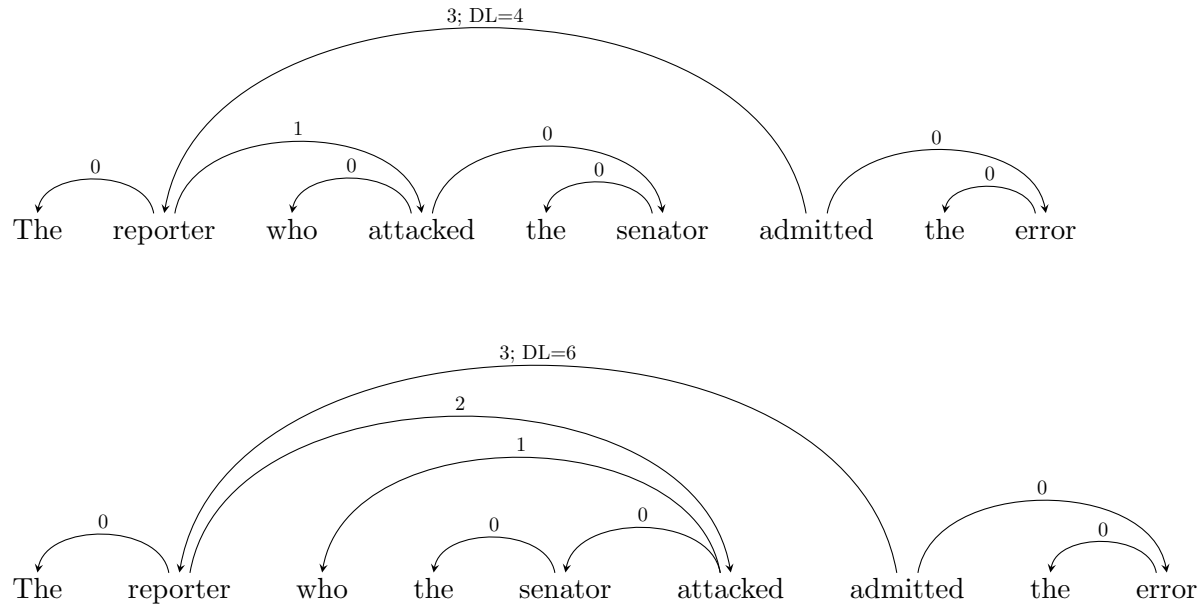


Figure 2: Lower overall dependency length (DL) of subject-relative (top) compared to object relative clause center-embeddings (bottom)

the amount of resources available to process language (Chomsky & Miller 1963, Schuler, AbdelRahman, Miller, & Schwartz 2010, Yngve 1960). Memory load can be estimated with EMBEDDING DEPTH (Wu et al. 2010), which captures the influence of the number of center embeddings (syntactic left branches within right branches).[3] Whereas DLT and embedding depth rely on the noisiness and effort of memory operations, SURPRISAL is a theory of neural activation allocation which quantifies the predictability of a given word in a syntactic or lexical context (Levy 2008).

2.1 Dependency Locality Theory

According to Gibson's (2000) Dependency Locality Theory (DLT), the syntactic complexity of a sentence is the sum of two kinds of processing costs, namely its STORAGE COST and INTEGRATION COST. Storage cost refers to the cost of maintaining in memory the syntactic predictions or requirements

[3]For a discussion of embedding depth as part of the prediction process, please refer to Linzen and Jaeger (2014; 2015).

of previous words. Integration cost is the cost of syntactically connecting a word to previous words with which it has dependent relations. The integration cost for a word increases with the distance to the previous words with which it is connected, on the grounds that the activation of words decays as they recede in time, making integration more difficult. Distance in Gibson's theory is measured in terms of the nature and the number of intervening discourse referents. Using self-paced reading experiments, Gibson demonstrated the greater processing complexity of object-extracted relative clauses (2-b) compared to subject-extracted relative clauses (2-a):

(2) a. The reporter who attacked the senator admitted the error
b. The reporter who the senator attacked admitted the error

Figure 2 depicts trees representing the above examples where dependency length is measured using intervening nouns and verbs as per Gibson's original definition. DLT predicts the tendency of human processing to prefer shorter dependencies in order to facilitate comprehension. While DLT predictions have been validated with eye-tracking data (Boston et al. 2008, Demberg & Keller 2008, Smith & Levy 2013), such studies have had difficulty observing the expected correlation with comprehension difficulty, and when they have observed a correlation in the correct direction (with longer dependencies inducing slower reading times), the predicted effect has been limited to rather long dependencies.

Extending DLT beyond language comprehension, Temperley (2007) poses the question: Does language production reflect a preference for shorter dependencies in order to facilitate comprehension? By means of a study of Penn Treebank data, Temperley shows that English sentences do display a tendency to minimize the sum of all their head-dependent distances. In phenomena involving syntactic choice, the tendency to minimize the overall dependency length is illustrated by facts like the greater length of subject noun phrases in inverted versus uninverted quotation constructions, greater length of postmodifying versus premodifying adverbial clauses, the tendency towards short-long ordering of postmodifying adjuncts and shorter length of the first adjunct compared to the second adjunct in clauses with three postmodifying adjuncts (these phenomena are illustrated using examples later in Section 4). Additionally, for head-final languages, dependency length minimization results in preverbal "long-short" constituent ordering in language production as evinced from studies on Japanese (Yamashita & Chang 2001), Korean (Choi 2007) and Basque (Ros, Santesteban, Fuku-

mura, & Laka 2015).[4]Gildea and Temperley (2010) report results from a tree-linearizing experiment, where given a dependency tree representation of an English sentence, the task is to order the children of each node using different methods. They investigate the problem of constructing a grammatical sentence using dynamic programming algorithms on projective tree structures to determine the word order of descendants of tree nodes. The algorithms (described in Gildea and Temperley 2007) order constituents based on the principle of minimizing dependency length and compare the dependency length of the output with that of actual English. Their results indicate that random linearizations have higher dependency lengths compared to English, while a dependency length based algorithm produces linearizations closer to actual English. Futrell, Mahowald, and Gibson (2015) extend these results by conducting a large-scale study of dependency length minimization involving 37 languages (see also Ferrer i Cancho 2004, Gulordava & Merlo 2015, Liu 2008). Futrell and colleagues demonstrate that for all the languages which were part of the study, the overall dependency length of a given natural language is shorter than the average of the artificially created baseline languages having no preference for dependency length minimization. But as all these authors note, dependency length minimization is only a tendency to be balanced by other factors, and a weaker one in freer word order languages like German.

Tily (2010) provides evidence that the pressure to minimize dependency length is significant in language change. Tily analyzes the diachronic trend towards dependency length minimization starting from Old English and moving towards Middle and Modern English. In Old English (OE) and Middle English (ME), both SVO and SOV orders were available and subjects as well as other preverbal dependents (including objects) were common. The study illustrates the tendency to avoid long dependencies between the verb and subject or other preverbal material by resorting to strategies like placing longer objects after the verb, thus ultimately leading to the frequent SVO order seen in Modern English.

2.2 Surprisal

Surprisal is an information theoretic characterization of comprehension difficulty expressed in bits, where lower values indicate lower processing load (Hale

[4]For a survey of the literature on cross-linguistic language production, see Jaeger and Norcliffe (2009).

2001, Levy 2008). More predictable words are associated with lower surprisal values in comparison to less predictable words. More predictable words are also known to induce faster processing times in reading (Boston et al. 2008, Demberg & Keller 2008). Mathematically, surprisal for word $k+1$ is defined using the conditional probability of a word given its sentential context. Mathematically, $S_{k+1} = -\log P(w_{k+1}|w_1...w_k)$. Practically, this is estimated using either simple lexical models like n-gram models or syntax-based, Probabilistic Context Free Grammars (PCFGs). Assuming strings of a language are generated by PCFG rules, the prefix probability of each word w_k is calculated by summing the probabilities of all trees T spanning words w_1 to w_k:

$$P(w_1...w_k) = \sum_T P(T, w_1...w_k) \tag{1}$$

Surprisal (Hale 2001) is then estimated by substituting this into the previous equation:

$$S_{k+1} = -\log \frac{P(w_1...w_{k+1})}{P(w_1...w_k)} = \log \sum_T P(T, w_1...w_k) - \log \sum_T P(T, w_1...w_{k+1}) \tag{2}$$

In addition to locality effects, anti-locality effects have been discussed in the sentence comprehension literature on German (Konieczny 2000) and Hindi (Vasishth & Lewis 2006). Like DLT, surprisal theory also predicts that object relative clauses have higher surprisal values compared to subject relative clauses and hence are harder to process. But for relative clauses, these two theories differ crucially in the actual word in the sentence where the processing difficulty occurs, with the DLT estimate being closer to observations (Levy 2008). They also make opposite predictions in the case of the verbal dependents in verb-final contexts in languages like Hindi and German. In this case, DLT predicts greater comprehension difficulty when a verb has more dependents since the cost of integrating more dependents is higher. However, experiments indicate a speed-up in reading times at the verb in cases where it has many dependents (Konieczny 2000, Vasishth & Lewis 2006). This is an effect predicted by surprisal theory. According to surprisal theory, a greater number of preverbal dependents provides greater syntactic context to the comprehender and hence sharpens the expectation about the location, nature and identity of the verb, which comes at the end, thus facilitating comprehension (Levy 2008).

Hawkins (2011) discusses anti-locality effects in detail in the context of relative clauses in German, contrasting relative clauses adjacent to their nominal heads with extraposed relative clauses. Though corpus counts and offline sentence judgement ratings preferred structures predicted by global measures of locality like Early Immediate Constituents (Hawkins 1994), online measures of comprehension like reading times did not reflect slowdowns predicted by locality. For example, Konieczny's (2000) paper reported faster reading times at the clause-final matrix verb with increasing head-dependent distance. Hawkins points to the possibility that ease of comprehension at certain points in a clause (the clause final matrix verb in Konieczny's study) might be offset by comprehension difficulty at earlier points in the same clause. Following this discussion, we use the term "non-locality" to refer to cases where locality constraints are not respected.

Surprisal theory also models other established findings in the literature for which other explanations had been proposed (Levy 2008, van Schijndel et al. 2014). Examples include English relative clause processing (MacDonald, Pearlmutter, & Seidenberg 1994, Traxler, Pickering, & Clifton 1998, J. Trueswell, Tanenhaus, & Garnsey 1994) as well as subject preference in disambiguating agreement and case marking conditions (Bornkessel, Schlesewsky, & Friederici 2002) and predictions of verbal subcategorization preference (Pickering & Traxler 2003, J. C. Trueswell, Tanenhaus, & Kello 1993).[5]

Since in this work we compare a complete corpus sentence to a constructed grammatical alternative, we use measures defined at the sentence level. The specific information theoretic measures we use are as follows:

1. n-GRAM LOG PROBABILITY for each word in a sentence is estimated using a 5-gram language model derived from the English Gigaword corpus (Parker, Graff, Kong, Chen, & Maeda 2011), a resource used widely in many mainstream Natural Language Processing (NLP) applications. It contains nearly 10 million documents with a total of around 4 billion words. The language model, based on a true-cased and PTB-tokenized version of the corpus, uses the KENLM[6] implementation of modified Kneser-Ney smoothing (James 2000) and is provided as part of the OpenCCG[7] NLP library. Individual per-word log probability values are summed to calculate the n-gram log probability for

[5]For an extensive review (and references therein) of prediction in language comprehension, see (Kuperberg & Jaeger 2016); for a more concise summary, see (Jaeger & Tily 2011)

[6]`http://kheafield.com/code/kenlm/`

[7]`http://openccg.sourceforge.net/`

the entire sentence. (The negative of this quantity gives total n-gram surprisal.) The Gigaword corpus in conjunction with modified Kneser-Ney smoothing is a state-of-the-art computational method which has been shown to be useful in NLP applications like machine translation and Natural Language Generation (NLG).

2. LATENT-VARIABLE PCFG LOG LIKELIHOOD of a sentence is estimated using a latent-variable PCFG parser which produces state-of-the-art parsing performance (Petrov, Barrett, Thibaux, & Klein 2006).[8] The likelihood of a sentence is calculated by summing the probabilities of all parse trees for the sentence. (Again, the negative of the PCFG log likelihood gives cumulative surprisal, this time based on a latent -variable PCFG.) In this work, the parser used a grammar based on standard WSJ training sections 02–21 to parse the Brown corpus and WSJ sections 00,01, 22, 23 and 24. WSJ sections 02-21 were parsed using jack-knifed grammars (trained by excluding any given test section) in order to prevent structural decisions being memorized because of the overlap between training and test sections.

   The grammar used by the parser in this work is inferred from the data by means of hierarchically state-split PCFGs using Petrov et al.'s split-merge latent-variable technique. Similar to distributional clustering of words, this latent-variable induction infers special categories from the context in which words occur. These categories capture more fine-grained syntactic and semantic distinctions than those in the original Penn Treebank, while they are not as specific as words. Petrov et al. describe many such patterns, for example the fact that verbs of communication such as *says* and *adds* are tagged using the same tag VBZ-4, while the tag VBZ-5 consists of verbs denoting propositional attitudes like *believes*, *means* and *thinks*. Similarly, phrasal rules are also split along the lines of root vs. embedded sentential contexts or finite vs. infinite verbal contexts.

2.3 Other Complexity Measures

In addition to dependency length, EMBEDDING DEPTH is also known to create processing difficulty (Chomsky & Miller 1963, Wu et al. 2010, Yngve 1960). Such effects could also be responsible for alternation choices during language production, so we test a variety of complexity measures

[8] The parser, popularly known as the Berkeley parser, is downloadable via `https://github.com/slavpetrov/berkeleyparser/`.

based on embedding depth in addition to our dependency length predictors. To calculate these complexity measures, an incremental probabilistic left-corner parser (van Schijndel, Exley, & Schuler 2013) based on the Petrov et al. (2006) latent-variable PCFG computes the n-best parses at each word, where n is the desired beam width. In this work, we have chosen a beam width of 3000, which was shown to be effective in pilot studies. Each parse is associated with its incremental likelihood given each successive lexical observation. The parser associates each syntactic node in each hypothesis with its embedding depth, weighted by the prefix probability of that hypothesis. The embedding depth of a parse increases whenever a non-terminal left branch in the syntax tree is generated from a right branch.

Weighted embedding depth increases the cost of maintaining increasing numbers of disjoint parse elements (Gibson 2000, Lewis, Vasishth, & Van Dyke 2006).[9] The more likely a parse hypothesis is, the more cognitive resources will be allocated to that hypothesis, which should increase the amount of cognition affected by that maintenance effort.[10] In the following evaluations, weighted embedding depth is computed as follows:

- A lexical item at position k is given a complexity score based on its embedding depth multiplied by its parse likelihood, which is summed over the set of active parse trees (T_k).

$$\textit{weighted embedding depth}_k = \sum_{t \in T_k} P_t(w_k \mid w_0 \ldots w_{k-1}) \cdot \text{depth}_t(w_k) \tag{3}$$

- The resulting scores are summed over the sentence (S).

$$\textit{weighted embedding depth} = \sum_{k \in S} \textit{weighted embedding depth}_k \tag{4}$$

Similarly, many psycholinguistic theories hypothesize that modifying embedding depth in working memory becomes harder as more elements are

[9] In Gibson (2000), this notion is indirectly reflected in the storage cost measure. The cost of performing a storage operation is dependent on the number of predictions that must be concurrently maintained.

[10] The use of a probability-weighted depth here assumes that alternative analyses of various depths are superposed in a distributed representation of attentional focus (Schuler 2014), rather than occupying single-element buffer-like memory slots, consistent with the idea that surprisal represents renormalization of superposed activation patterns to a constant magnitude after analyses that are inconsistent with observed words have been filtered out.

parse trees	$k = 1$	$k = 2$	$k = 3$
t_1	$P(w_1^{d2} \mid w_0^{d1}) = 0.2$	$P(w_2^{d2} \mid w_0^{d1}\ w_1^{d2}) = 0.2$	$P(w_3^{d1} \mid w_0^{d1}\ w_1^{d2}\ w_2^{d2}) = 0.7$
t_2	$P(w_1^{d2} \mid w_0^{d1}) = 0.4$	$P(w_2^{d3} \mid w_0^{d1}\ w_1^{d2}) = 0.3$	$P(w_3^{d2} \mid w_0^{d1}\ w_1^{d2}\ w_2^{d3}) = 0.2$
t_3	$P(w_1^{d2} \mid w_0^{d1}) = 0.4$	$P(w_2^{d3} \mid w_0^{d1}\ w_1^{d2}) = 0.5$	$P(w_3^{d3} \mid w_0^{d1}\ w_1^{d2}\ w_2^{d3}) = 0.1$
weighted embedding depth$_k$	$2 \cdot (0.2 + 0.4 + 0.4) = 2$	$2 \cdot 0.2 + 3 \cdot (0.3 + 0.5) = 2.8$	$1 \cdot 0.7 + 2 \cdot 0.2 + 3 \cdot 0.1 = 1.4$
weighted embedding depth	$(1) + 2 + 2.8 + 1.4 = 7.2$		
1-best embedding depth	$(1) + 2 + 2 + 1 = 6$		

Table 1: Incremental parser beam examples and associated hypothesis likelihoods for the sequence $w_0\ w_1\ w_2\ w_3$ (upper section). Superscripts denote the syntactic embedding depth of each word. Each column denotes another time step k of the parse. For example, at time step one ($k = 1$), there are three partial parses with normalized probabilities 0.2, 0.4 and 0.4 (resp.), all of which extend to nesting depth 2, while at time step two ($k = 2$), parse t_1 remains at depth 2 but parses t_2 and t_3 extend to depth 3. Incremental complexity measures (middle section) are summed over each sentence to give the ultimate measures used in the evaluation (lower section). The calculations of *embdif*$_1$ and *1-best embedding depth* presume $k = 0$ had a weighted embedding depth of 1, which is reasonable when starting a new sentence at w_0.

stored in working memory (Gibson 2000, Lewis et al. 2006, Schuler 2014, van Schijndel et al. 2013). Finally, parsing may occur serially (i.e. only a single hypothesis may be considered at a time), or the best (intended) parse may be the only hypothesis that exerts a measurable influence during sentence generation. To capture this notion, we use a measure of lexical depth (*1-best embedding depth*), which we compute by summing the embedding depths from the most probable final parse T given the entire, non-incremental observation sequence:

$$\textit{1-best embedding depth} = \sum_{w \in T} \text{depth}_T(w) \tag{5}$$

To illustrate, consider the incremental parse hypotheses in Table 1. For each time step, the complexity measures are given. Note that *1-best embedding depth* is not computed incrementally; instead, the embedding depths of each observation in the best scoring parse are summed after the parse is complete.

3 Other Factors Influencing Constituent Ordering

This section discusses other factors which have been described in the literature as influencing constituent ordering. We discuss the relationship of these factors with constituent length and dependency length minimization. As K. Bock, Irwin, and Davidson (2004) discuss, the factors affecting constituent order can be divided into two main groups: (i) elemental factors operating at the level of elements of an utterance (words); and (ii) structural factors operating at the level of syntactic structure.

According to some previous theories of language production, cognitive accessibility is the single most important factor that governs elemental processes in constituent ordering. More accessible elements are produced first in comparison to less accessible elements which are realized subsequently (Arnold 2008, J. K. Bock 1982, Ferreira & Dell 2000). Alignment-based accounts of production (J. K. Bock & Warren 1985) propose that grammatical function assignment is aligned with the relative accessibility of elements. Here accessibility is conceived as the *conceptual accessibility* of elements. Conceptual accessibility is predicated upon inherent features like animacy, imageability or prior discourse mention. However, availability-based accounts of language production (V. Ferreira 1996, Ferreira & Dell 2000), in addition to inherent properties mentioned above, consider accessibility effects to be more direct. Here, accessibility is the ease with which linguistic elements are retrieved from memory. Previous studies have shown that accessibility is influenced by the following factors:

1. **Animacy**: Animate nouns tend to precede inanimate nouns since they are more accessible (J. K. Bock & Warren 1985) and this is independent of length in influencing constituent ordering choices (Snider & Zaenen 2006). Snider and Zaenen analyze the effect of animacy on NP fronting and the interaction between animacy and heaviness. They conclude that inanimate entities are more likely to occupy the topic position while animate entities are more likely to be left-dislocated. Heavier constituents are likely to be topicalized or left-dislocated compared to light ones, going against purely linear order based accounts. Overall, their study put forth the view that animacy and length independently influence ordering choices. Such effects can potentially be modelled using PCFG surprisal estimated from a sufficiently large corpus of the language with fine-grained lexical categories encoding animacy.

2. **Information status considerations**: Given elements (either men-

tioned in the prior discourse or part of the context) tend to precede new elements. As Arnold (2011) notes, previous studies have indicated that first person pronouns like *I* and *we* are very accessible compared to definite NPs (*the cyclist*, for example). In contrast, indefinites like *a cyclist* are much less accessible. Correlations between discourse status and length have been noted in the literature. For example, Arnold notes that the first mention of a new discourse referent tends to be a long NP (*The avid cyclist who also teaches linguistics*), but subsequently, this is given information resulting in the use of shorter expressions like the pronoun *he/she*. Arnold et al. (2000) tested the effect of heaviness and newness of constituents in determining constituent order choices using a corpus study as well as a production experiment. Both length of NPs and discourse status (whether an element is given or new) contribute towards constituent ordering in the case of dative alternation and heavy noun phrase shift. Though both relative length of constituents and discourse status were significant predictors of order, heaviness accounted for more of the variation compared to discourse status. Discourse newness has an effect when heaviness does not make any predictions in either direction. In their study of dative alternation, Bresnan et al. (2007) reported both these factors to be independent predictors of the choice of the dative realization. These studies point to the conclusion that discourse status is a factor which is independent of the drive to minimize dependency length and it needs to be considered separately when deciding between competing ordering options (Gallo, Jaeger, & Smyth 2008, Snider 2009).

Currently, our model of surprisal does not go beyond the sentence level. Thus information status considerations going beyond the lexical or clausal level are not modelled by surprisal. However in future work, surprisal can be linked to information status considerations by linking it to predictability across discourse units extending beyond the sentence. Qian and Jaeger (2012) develop a quantitative model of exponential cue decay across discourse units spanning multiple sentences and validate it using data from 12 languages. This framework can be augmented to estimate the givenness and newness of a given discourse referent. Thus given elements would be more predictable (low surprisal value) in contrast to new elements (high surprisal value).

3. **Semantic connectedness**: Another factor which the literature discusses is the semantic connectedness between the verb and its dependent constituents. Wasow and Arnold (2003) discuss cases involving

idioms (e.g., *take our concerns into account*) and collocations (e.g., *bring that debate to an end*). They report that 26% of non-idiom examples were in the non-canonical shifted order while around 60% of the idioms displayed shifting. Hawkins (2001) also studied the role of meaning in constituent ordering. Length can override semantic connectedness of verb and postverbal constituents. He examined postverbal prepositional phrases and reported that constituents with a greater semantic degree of connectedness with the verbal head (ascertained using entailment tests) occur more adjacent to the verb. In Hawkins' framework, which relies on constituency representations of syntax, semantic connectedness sets up additional dependencies between words in addition to their syntactic sisterhood within phrases, and and thus enhances the preference for locality (Hawkins 2004). The cited work shows that such additional dependencies do indeed result in tighter adjacency or locality compared with less dependent controls.

The following structural factors have been discussed in previous work:

1. **Construction type**: Syntactic priming experiments suggest that speakers tend to use certain constructions like active voice (over passive voice). Speakers are also prone to repeat structures used by interlocutors in the preceding discourse (J. K. Bock 1986, W. Levelt & Maasen 1981, Pickering & Branigan 1998). This is independent of length. In this work, we analyze different construction types and surprisal integrates lexical cues about constructions.

2. **Syntactic complexity**: Syntactic complexity and length are factors which independently influence constituent ordering in many constructions (Wasow 2002, Wasow & Arnold 2003). Following Chomsky and Miller's (1963) original intuition that syntactic complexity could have an effect on the processing of syntactic structures independent of length, Wasow and Arnold (2003) examine the effect of these factors in conjunction as well as in isolation. Here it should be noted that their definition of complexity is the presence of a clause. To test the relationship between length and complexity they conducted a questionnaire study where subjects were asked to assign acceptability judgements to stimuli containing both complex and simple NPs (controlled for length) in both shifted as well as unshifted positions as shown below. They examined the following constructions: Heavy Noun Phrase Shift (HNPS), dative alternation and the verb-particle construction. The

following examples from the paper illustrate the types of stimuli used (emboldened words have dependencies with the verb *took*):

(3) a. John **took** only the **people** he knew into account. [Unshifted]
b. John **took into** account only the **people** he knew. [Shifted]
c. John **took** only his own personal **acquaintances into** account. [Unshifted]
d. John **took into** account only his own personal **acquaintances**. [Shifted]

The results suggest that when total length is controlled, syntactic complexity independently contributes to ordering preferences. Thus complexity is a factor which might have a bearing on the choice between two constituent orders with equal dependency lengths. To test the effect of these factors when both of them vary, they conducted a corpus study based on the aligned Hansard corpus and examined the number of words and syntactic complexity in the constructions mentioned above. When both length and syntactic complexity vary, both length and syntactic complexity are significant predictors of ordering independent of each other in the case of HNPS and dative alternation. Moreover, in the case of constituent length, the relative length of the constituents determines ordering choices rather than the length of either one alone. But for the verb-particle construction, length significantly contributes to ordering, while syntactic complexity does not seem to have much of an effect: since the particle is a light constituent, sentences with object noun phrases greater than three words always display the joined verb-particle pattern irrespective of syntactic complexity. This work also confirms the tendency for short-long constituent orders that had previously been reported in the literature (in form of proposals like the *principle of end weight*). Thus for HNPS and dative alternation, dependency length minimization is not the only driver of production: syntactic complexity (defined as the presence of a clause) also independently influences production choices. Embedding depth measures model syntactic complexity in our study.

3. **Lexical bias**: The verb influences the choice of realization in dative alternation (Bresnan et al. 2007, Gries 2005, Gries & Stefanowitsch 2004, Wasow & Arnold 2003) and can also influence phenomena like heavy NP shift (Stallings, MacDonald, & O'Seaghdha 1998, Staub,

Clifton, & Frazier 2006) and passivization (Manning 2003). Dative alternation is influenced by the verb as certain verbs have a bias towards the choice of the realized dative (Bresnan et al. 2007, Wasow & Arnold 2003). Anttila, Adams, and Speriosu (2010) extend this proposal by examining the difference between one foot and two foot verbs in dative alternation. They show that the PP-choice in dative alternation and HNPS is more common with two-foot verbs. Thus if rhythmic feet in words are counted as part of dependency length calculations (in a revised definition), this factor is directly related to dependency length minimization. Further, in the case of heavy NP shift, both comprehension (Staub et al. 2006, van Schijndel et al. 2014) and production (Stallings et al. 1998) studies have shown that the properties of individual verbs (e.g., transitivity) can influence the shifting of NPs. This has a direct effect on dependency length calculations, and thus this factor does interact with the minimal dependency length preference. Syntactic surprisal models lexical bias by incorporating categories reflecting the properties of different verbs like transitivity.

4. **Prosodic factors**: The *principle of end-weight* stipulates that longer or heavier constituents tend to come later in the clause. In the literature, weight has been calculated in terms of words or syntactic nodes (Wasow 2002), but Anttila et al. (2010) derive end-weight effects from stress and prosodic units in an Optimality Theory (OT)-based constraint ranking framework. In an experiment which correlates eight different measures of weight with responses in dative alternations (i.e. *NP* vs. *PP* realization), they show that the log number of primary stresses in the theme shows the greatest correlation with the correct response. This finding has the consequence that lexically unstressed words like function words (*the, a*, for example) do not contribute towards weight. This has implications for the calculation of dependency length, as discussed previously. Lee and Gibbons (2007) also provides experimental evidence of stress-based optimization in speech production. In this work, we do not model prosodic stress.

5. **Complement-Adjunct distinction**: Hawkins (2001) argues that complements lie closer to the verbal head because of the presence of more combinatory or dependency links between complements and heads. Lohse et al. (2004) provided corpus-based evidence for this proposal. In this work, dependency links were detected using entailment tests of the form: "Does $V\ PP_1\ PP_2$ entail V alone or does V have

	Frequency	Mean Length	Mean Distance to verb head
Adjunct	1326	6.11	4.57
Argument	4266	7.70	2.26

Table 2: Postverbal argument-adjunct patterns in PTB Sect00 data using Propbank annotation

a meaning dependent on either PP_1 or PP_2?" This is exemplified by the following sentences:

(4) a. The man waited for his son in the early morning
b. The man waited
c. The man counted on his son in his old age
d. The man counted

Example (4-a) above entails (4-b), but (4-c) does not entail (4-d). One other reason why complements tend to be adjacent to their verbal heads is that complements, unlike adjuncts, are specified in the lexical co-occurrence frame of the head (Pollard & Sag 1994). Thus complements, which are more central to the meaning of the sentence, display a tendency to be closer to the verbal head. This preference often results in overriding the preference for minimizing dependency length.

Following Hawkins' work, we also conducted a preliminary investigation of the relationship between arguments and adjuncts and their respective verbal heads in the Penn Treebank data. The complement-adjunct distinction was obtained from Propbank roles (Palmer, Gildea, & Kingsbury 2005), a set of manually annotated verbal semantic roles. Postverbal distances were calculated by counting the number of words separating the head and the left edge of constituents. Table 2 illustrates these results. It can be seen that postverbal arguments are closer to verbal heads compared to postverbal adjuncts, confirming the patterns observed in Hawkins' study.

Using a grammar with correct distinctions can enable surprisal to quantify the argument-adjunct distinction and thus model semantic connectedness as described in Hawkins (2004). As Levy (2008) discusses, the structure of PCFGs can incorporate morpho-syntactic

properties like case marking and agreement in addition to unbounded dependencies like relativization into syntactic categories. A given category may be conditioned on the lexico-semantic contents of its governor. Local domains are modelled using history-based conditioning on sister nodes. At the same time, the probability of a given node can also be conditioned on its grandparent and sisters of the grandparent.

In related work, Wiechmann and Lohmann (2013) quantify the relative impact of various factors on the ordering of English postverbal PP phrases. They considered factors like semantic connectedness, syntactic weight, functional generalizations like Manner-Place-Time (MPT) order of adjuncts and pragmatic differences in information structure. They showed that syntactic weight minimization accounted for most of the data, but at the same time, the magnitude of semantic connectedness was greater compared to syntactic weight. Thus semantic connectedness predicted PP orders correctly when weight is pulling in the opposite direction. The contributions of the MPT generalization and pragmatic information status, though statistically significant, only led to small increases in classification accuracy while predicting the corpus choice.

4 Data

As noted in the introduction, the datasets used in the study are the Brown (Francis & Kučera 1989) and Wall Street Journal (WSJ) portions of the Penn Treebank (PTB) corpus (Marcus et al. 1993), a standard resource for natural language processing applications. Both corpora contain syntactically annotated written text from various domains and genres. WSJ contains newswire text while the Brown corpus contains sentences from around 15 genres of American English text published in 1961. From the constituent structure syntax trees provided in these corpora, we extracted the subset of constructions involving syntactic choice in Temperley's (2007) earlier study.[11] In addition, we also extracted dative alternation cases.[12] Table 3 shows the frequency of the syntactic choice constructions used in our study. Properties of the domain of each dataset is also visible there. The WSJ corpus is primarily journalistic text where inverted quotation constructions are much

[11] The syntactic choice constructions were extracted using the *tgrep* patterns provided in the appendix of (Temperley 2007).

[12] For the dative alternation construction, we used the same list of verbs created by Bresnan et al. (2007) available via the languageR package in R.

Construction	Subtype (Frequency)	Frequency
Dative alternation	NP-PP (33; 344) NP-NP (505; 799)	538; 1143
Quotation	Inverted (54; 1764) Uninverted (549; 2301)	603; 4065
Postverbal adjuncts	1-constituent (2213; 4366) 2-constituents (2259; 4539) 3-constituents (1116; 3061)	5588; 11966
Preverbal adjuncts	1-constituent (1401; 2483) 2-constituents (255; 673)	1656; 3156

Table 3: Frequency of syntactic choice constructions in the (Brown; WSJ) corpora

more frequent compared to Brown corpus text comprising of text from multiple genres.

Subsequently, we created syntactic variants by manipulating the extracted trees. For this purpose, we used hand-crafted rules over gold standard trees. So the variants are all expected to be high quality. The following subsections exemplify the constructions and their subtypes. In each example group, the first sentence is the Brown corpus sentence followed by hand-crafted variants.

4.1 Dative alternation

A reference sentence with NP-NP structure is transformed into the NP-PP variant:

(5) a. Just about the most enthralling real-life example of meeting cute is the Charles MacArthur-Helen Hayes saga: reputedly all he did was give [her] [a handful of peanuts], but he said simultaneously, "I wish they were emeralds." (CF01.2)

b. Just about the most enthralling real-life example of meeting cute is the Charles MacArthur-Helen Hayes saga: reputedly all he did

was give [a handful of peanuts] [to her], but he said simultaneously, "I wish they were emeralds."

A reference sentence with the NP-PP structure is transformed into the NP-NP variant:

(6) a. "Our information is that she gave [the proceeds of her acts] [to Jelke]." (CF09.23)
b. "Our information is that she gave [Jelke] [the proceeds of her acts]."

4.2 Quotations

A V-S reference sentence structure is transformed into a variant with S-V structure:

(7) a. "Hang this around your neck or attach it to other parts of your anatomy, and its rays will cure any disease you have," said [the company]. (CF10.75)
b. "Hang this around your neck or attach it to other parts of your anatomy, and its rays will cure any disease you have," [the company] said.

Similarly, reference sentences with uninverted quotations are transformed into variants with inverted V-S structure:

(8) a. "It's people of your own kind," a girl remarked. (CF25.67)
b. "It's people of your own kind," remarked a girl.

4.3 Postverbal Adjuncts

For sentences containing one postverbal adjunct, a variant is created by placing it before the clause it modified:

(9) a. Hardly anyone ashore marked her [as she anchored stern-to off Berth 29 on the mole]. (CF02.4)
b. [As she anchored stern-to off Berth 29 on the mole], hardly anyone ashore marked her.

For reference sentences with two postverbal adjuncts, one other variant is created by interchanging these adjuncts:

(10) a. It had been made shockingly evident [that very morning] [to Ensign Kay K. Vesole, in charge of the armed guard aboard the John Bascom]. (CF02.49)
b. It had been made shockingly evident [to Ensign Kay K. Vesole, in charge of the armed guard aboard the John Bascom] [that very morning].

Only these two variants were considered as in Temperley's study. For reference sentences with three postverbal adjuncts, five other variants are created by permuting these adjuncts:

(11) a. Oranges and grapefruit are shipped [from Florida] [weekly] [from an organic farm]. (CF04.86)
b. Oranges and grapefruit are shipped [weekly] [from Florida] [from an organic farm].
c. Oranges and grapefruit are shipped [from an organic farm] [weekly] [from Florida].
d. Oranges and grapefruit are shipped [from Florida] [from an organic farm] [weekly].
e. Oranges and grapefruit are shipped [from an organic farm] [from Florida] [weekly].
f. Oranges and grapefruit are shipped [weekly] [from an organic farm] [from Florida].

4.4 Preverbal Adjuncts

The variant corresponding to the reference containing one preverbal adjunct is created by post-posing the adjunct to after all VP constituents:

(12) a. [After the preliminary business affair was finished], Depew arose and delivered the convincing speech that clinched the nomination for Roosevelt. (CF03.67)
b. Depew arose and delivered the convincing speech that clinched the nomination for Roosevelt, [after the preliminary business affair was finished].

The variant corresponding to the reference sentence containing two preverbal adjuncts is created by interchanging the two:

(13) a. [In other words], [like automation machines designed to work in tandem], they shared the same programming, a mutual understanding not only of English words, but of the four stresses,

Label	Meaning
PCFG log likelihood	Sentence log likelihood emitted by a latent-variable parser (negative of this quantity gives cumulative PCFG surprisal)
ngram log likelihood	5-gram gigaword log probability (negative of this quantity gives n-gram surprisal)
weighted embedding depth	sum of beam embedding depths $\times$ parser probability
1-best embedding depth	Sum of embedding depths of non-punctuation lexical items in the best parse

Table 4: Glossary of terms

pitches, and junctures that can change their meaning from black to white. (CF01.7)

b. [Like automation machines designed to work in tandem], [in other words], they shared the same programming, a mutual understanding not only of English words, but of the four stresses, pitches, and junctures that can change their meaning from black to white.

Here again only these variants were considered as in Temperley's study.

5 Models and Results

This section describes the experiments we conducted and reports the main findings of this study.

Section 5.1 describes the model and experimental results investigating whether syntactic choice is influenced by dependency length amidst other controls. Section 5.2 explores the individual and relative contributions of the factors in predicting syntactic choice. Section 5.3 presents results of binning experiments which investigate the relationship between dependency locality and surprisal as a function of dependency length. Section 5.4 reports on the results of experiments involving constructions that aim to map the relative contribution of frequency and memory measures in predicting syntactic choice. A glossary providing the names and descriptions of the independent variables appears in Table 4.

5.1 Experiments with Regression Models

As mentioned in the introduction, we seek to extend previous work (Hawkins 1994; 2004, Temperley 2007) which has already established dependency length as individually influencing syntactic choice. Section 8.1 in Appendix A describes ranking experiments in which the relative merits of three distinct dependency length measures proposed in the literature are compared. Consequently, Gibson's definition of dependency length—measured by counting the number of discourse referents—is the measure we consider for all our subsequent experiments (referred to as *dependency length* from now on). Using Gibson's measure, we now investigate whether dependency length is a significant predictor of syntactic choice even when other cognitively grounded measures of comprehension are included as controls.

5.1.1 Ranking Model

Typically, both behavioural experiments (Arnold et al. 2000, Stallings et al. 1998, Staub et al. 2006) and corpus studies (Bresnan et al. 2007, Szmrecsanyi 2004, Wasow 2002) related to syntactic choice focus on a single or a very limited set of constructions. In contrast, our study is conceived as an investigation involving multiple construction types (other cross-construction studies include Reitter, Keller, & Moore 2011, Reitter & Moore 2014). Though Temperley (2007) considers multiple constructions, each construction in the corpus (e.g., postmodifying adverbial clauses) is directly compared to another construction with the opposite constituent ordering pattern (premodifying adverbial clauses, for example) and significance is reported by comparing average constituent lengths between the two constructions. In contrast, we generalize over all constructions by first creating plausible grammatical variants for all reference sentences in the Brown and WSJ corpora that exhibit syntactic choice phenomena discussed in Temperley's work (see examples in Section 4), then defining a ranking model that seeks to correctly rank order each pair of a reference sentence and a grammatical variant such that the reference sentence always outranks the variant.

Joachims (2002) shows how a SVM classifier can be used for ranking by classifying whether a pair of comparable items is in the correct rank order, which reduces to training a classifier on the difference of the feature vectors. We adapt this idea to a Generalized Linear Model (GLM) setting. For data involving categorical outcomes (binary in this case), GLMs are standard models designed to estimate the probability of outcomes using logistic regression. During training, maximum likelihood estimation is used

Data Point Label	Feature Vector	Feature Values		
		dependency length	PCFG log likelihood	ngram log likelihood
ref	Φ(ref)	30	-137.44	-59.44
var_1	$\Phi(\text{var}_1)$	30	-135.89	-61.16
var_2	$\Phi(\text{var}_2)$	32	-135.79	-58.09

(a) Original data points

Data Point Label	Condition	Feature Vector Difference	Feature Value Differences		
			dependency length	PCFG log likelihood	ngram log likelihood
1	$s_1 = \text{ref}$ $s_2 = \text{var}_1$	$\Phi(s_1) - \Phi(s_2)$	0	-1.55	1.72
0	$s_1 = \text{var}_2$ $s_2 = \text{ref}$	$\Phi(s_1) - \Phi(s_2)$	2	1.65	1.35

(b) Transformed data points

Table 5: Illustration of ranking model technique

to select model parameters, which in the case of GLMs involves iterative fitting techniques (Baayen 2008).

In the Joachims ranking setup, given any pair of comparable data points s_1 and s_2, $\Phi(s_1)$ and $\Phi(s_2)$ represent feature vectors encoding individual feature values (comprehension measures, in our case) of the data points. We train a logistic regression classifier on $\Phi(s_1) - \Phi(s_2)$, the difference in the feature vectors for all such points in the dataset. Half of the pairs are designated to have the reference sentence first ($s_1 = \text{ref}$) and the remaining half have the reference sentence second ($s_2 = \text{ref}$). Pairs where the reference sentence is correctly ordered first (i.e., where $s_1 = \text{ref}$) are coded as 1, with the rest coded as 0.

We illustrate how the dependent and independent variables are computed using the following examples involving one reference sentence and two syntactic variants:

(14) a. *Reference sentence* (ref): "One afternoon during a cold, powdery snowstorm, Fogg took off for Concord from the St. John field." (CF05.87.0)
b. *Variant 1*: "During a cold, powdery snowstorm one afternoon, Fogg took off for Concord from the St. John field."
c. *Variant 2*: "One afternoon during a cold, powdery snowstorm, Fogg took off from the St. John field for Concord."

Table 5 depicts the calculations for the above examples. Note that the use of relative values of features emerges naturally from viewing the task as a ranking task. This also confers the added benefit that feature values across sentences of varying lengths in the datasets are centered. Other possibilities such as using a binary dependent variable (say *early* vs. *late*) would only allow modelling 2 choices. In our dataset, we have cases involving choice of ordering 3 postverbal adjuncts leading to 3! possible variants. Thus we believe the method described above generalizes to any number of variants.

In their study of dative alternation, Bresnan and colleagues consider the dependent variable to be whether the recipient is expressed as a PP. Equivalently, they could have characterized this as the recipient realized late; or they could have coded this as theme being realized late (in which case the signs of all the predictors would have flipped). With our inverted subjects following quotations, we could code this as the subject being realized late. But it is unclear which one of late theme vs. late recipient would make sense together with late subject. A dilemma will also arise with preverbal and postverbal adjuncts, where it is less obvious how to identify each option. Using corpus choice as the dependent variable gives a common footing to what the independent variables are predicting.

Joachims (2002) shows that the learned model can be used for prediction by comparing the dot product of the learned feature weights (model parameters) $\mathbf{w}$ with the feature values for s_1 to the dot product of $\mathbf{w}$ with the feature values of s_2. In particular, s_1 is predicted to outrank s_2 when the dot product is greater,

$$\mathbf{w} \cdot \Phi(s_1) > \mathbf{w} \cdot \Phi(s_2) \tag{6}$$

or equivalently when the dot product with the feature difference is positive:

$$\mathbf{w} \cdot (\Phi(s_1) - \Phi(s_2)) > 0 \tag{7}$$

The same holds true in the logistic regression setting.

In order to investigate whether dependency length is a significant predictor of syntactic choice, we use the following GLM to predict whether s_1 is the corpus sentence in a pair (s_1, s_2):[13]

$$\begin{aligned} \textit{choice} \sim\ & \textit{PCFG log likelihood} + \textit{ngram log likelihood} + \textit{dependency length} \\ & + \textit{weighted embedding depth} + \textit{1-best embedding depth} \end{aligned} \tag{8}$$

Here the dependent variable *choice* is a binary choice variable where 1 denotes the correct choice and 0 stands for the incorrect choice. The independent variables are the measures of comprehension listed in Table 4.

5.1.2 Regression Results

The regression model results demonstrate that dependency length is a significant predictor of syntactic choice for both corpora (see Table 6). In fact, the table shows that all the independent variables used in the model are significant predictors of syntactic choice. The negative coefficient of the variable *dependency length* shows that relatively lower values of dependency length predict the corpus choice as opposed to the variant. Thus the tendency for dependency length minimization in written English attested in Temperley's (2007) corpus study is confirmed here even when other cognitively grounded controls are present. Latent-variable PCFG cumulative surprisal difference (negative of variable *PCFG log likelihood*) has a negative coefficient since *PCFG log likelihood* has a positive regression coefficient. This means that the corpus choice is predicted by relatively lower values of surprisal. Thus the results for surprisal (both PCFG and n-gram) are as one would expect, since these measures are based on models trained to maximize an objective based on the likelihood of the training data.

Moreover, trends of the regression coefficients are along the lines of results reported in the sentence comprehension literature, where low values of dependency length and surprisal are associated with ease of comprehension (Gibson 2000, Hale 2001, Levy 2008).[14] For the Brown corpus we also experimented with a Generalized Linear Mixed Model (GLMM) having genre (Brown corpus has text from 8 genres) as the random effects term.

[13] In this paper, models are presented in R GLM format where the dependent variable occurs to the left of '$\sim$' and independent variables occur to the right.

[14] For all the independent variables used in the study, we visually illustrate the relationship between their regression coefficients and the probability of predicting the correct choice by means of effects plots (see Figures A.2 and A.3 in Appendix A).

Predictor	Brown	WSJ
PCFG log likelihood	20.09, $p < 2e-16$	44.75, $p < 2e-16$
ngram log likelihood	28.96, $p < 2e-16$	39.73, $p < 2e-16$
dependency length	-15.90, $p < 2e-16$	-20.15, $p < 2e-16$
weighted embedding depth	-10.58, $p < 2e-16$	-10.89, $p < 2e-16$
1-best embedding depth	-3.35, $p = 0.00079$	-6.84, $p = 7.82e-12$

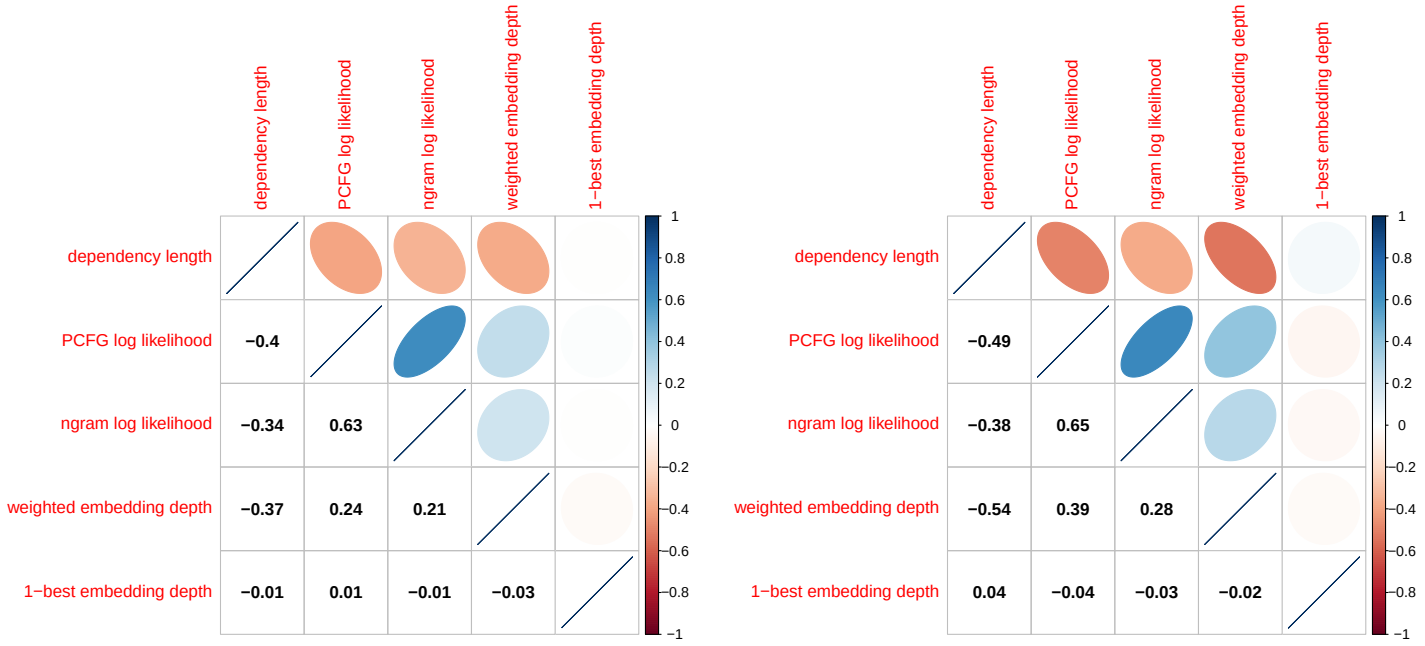
Table 6: Regression model testing the effect of predictors on syntactic choice using Brown (8385 data points) and WSJ corpora (20330 data points)

However, the latter model was not significantly different from the regression model discussed above. Thus genre is not a significant predictor of the corpus choices we investigate in this work.

Since all the independent variables used in the study emerged as significant predictors of syntactic choice, we calculated Pearson's coefficient of correlation between dependency length and the other independent variables. We found a low to moderate correlation between surprisal and dependency length values in our data (see Figure 3). In the case of the Brown corpus, dependency length exhibits low correlation with surprisal (and all other variables as well); in the WSJ corpus, dependency length correlates only moderately with surprisal. The variance inflation factors for each of the predictors are also in the reasonable range, with no one predictor conflated with the others. The low correlation between dependency length and surprisal has also been noted by Demberg and Keller (2008) for modelling reading times. Thus it is plausible that dependency length and surprisal are modelling different parts of the data, a conjecture which is borne out in our investigations described in a separate section, where we also present a comparison with Demberg and Keller's results.

5.2 Classification Experiments

This section explores the individual and relative contributions of the comprehension measures in predicting syntactic choice. To determine individual performance, each predictor is used to rank the reference sentence against each of the variants, with ties resolved by choosing one alternative randomly and then averaging results across 10 runs. For the Brown corpus, 5-gram gigaword surprisal (*ngram log likelihood*) is the most successful predictor, while for WSJ, surprisal based on the latent-variable parser (*PCFG log likelihood*) is the top predictor (Figure 4). For both corpora, dependency length (*dependency length*) is the next most effective predictor, with *weighted em-*



(a) Brown corpus

(b) WSJ corpus

Figure 3: Correlation plot of predictors

bedding depth also providing competitive performance.

To determine relative performance in prediction, each independent variable is successively added to the regression model. Across the two corpora, the latent-variable parser log likelihood, 5-gram log likelihood and dependency length produce significant improvements in classification accuracy over the ablated model not containing that particular factor (Figure 5). For each corpus, Likelihood Ratio Tests comparing models corresponding to successive bars of the bar plot also indicate that each model is significantly different from the ablated model shown in the previous bar.

5.3 Binning Experiments

In this section, we investigate in greater detail the relationship between dependency length and surprisal. As mentioned previously, these measures display only low to moderate correlation in our syntactic choice data. In

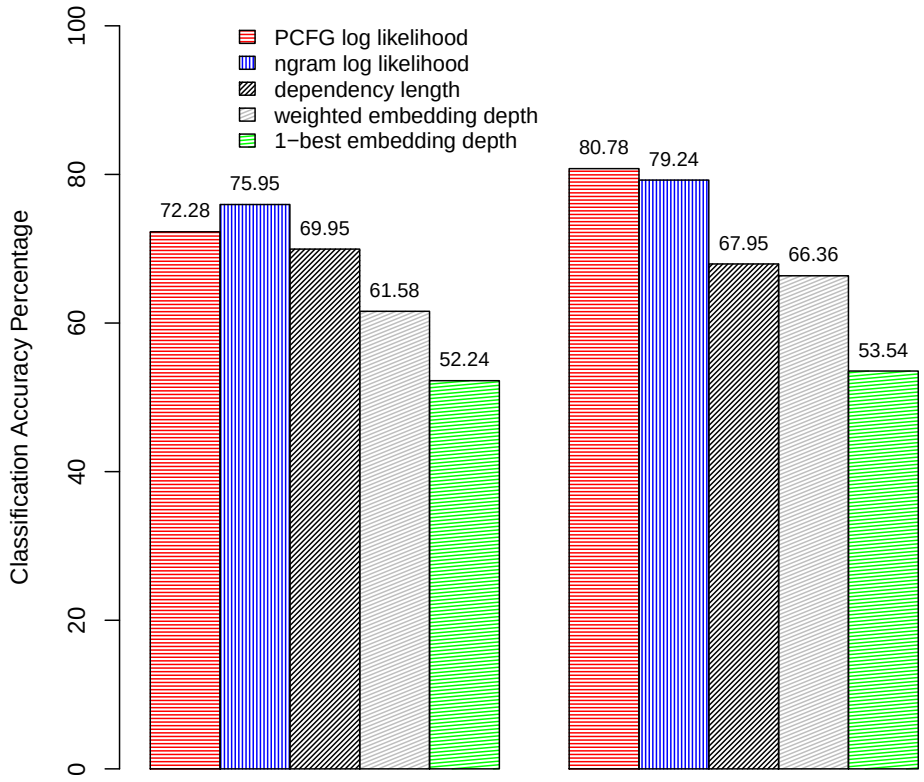


Figure 4: Classification accuracy of individual measures for the Brown (left) and WSJ (right) corpora

the context of sentence comprehension, Demberg and Keller (2008) report that surprisal and dependency length are not correlated and suggest they have complementary effects when predicting reading times in the Dundee corpus. However, they found that only large values of dependency length (integration cost) are effective for this task. We begin by examining the accuracy of dependency length and surprisal in predicting syntactic choice as a function of dependency length, then present a more detailed comparison with Demberg & Keller's results.

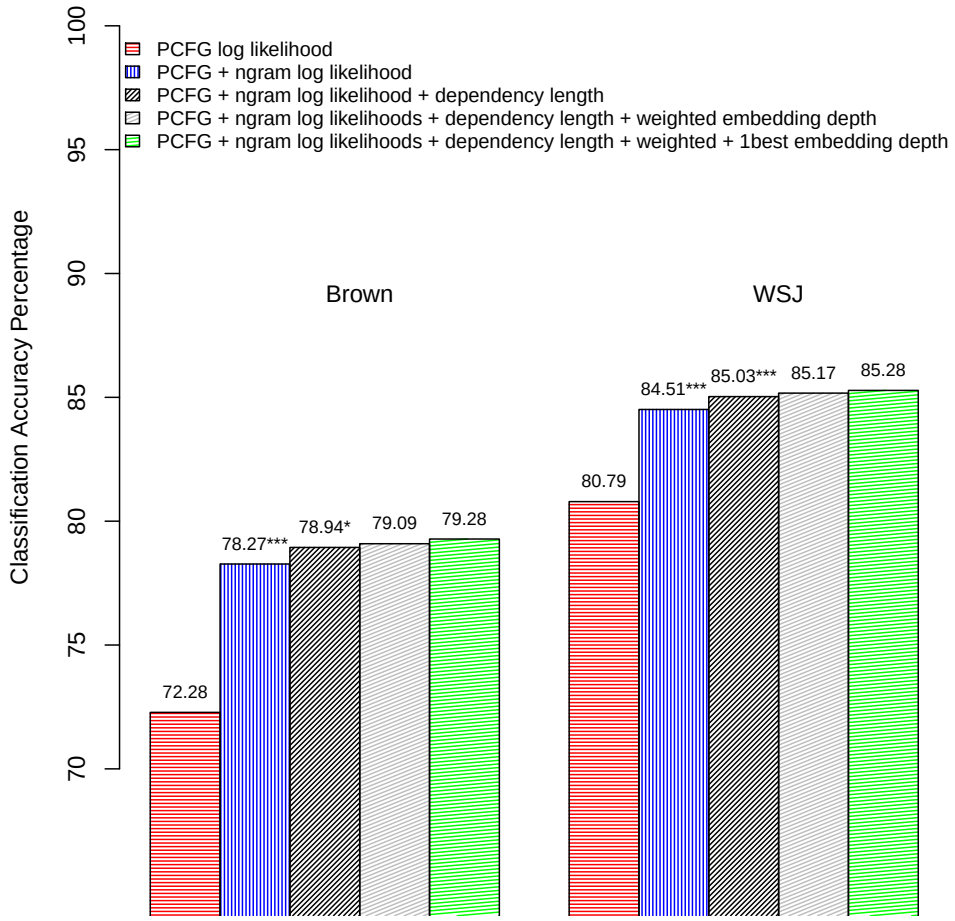
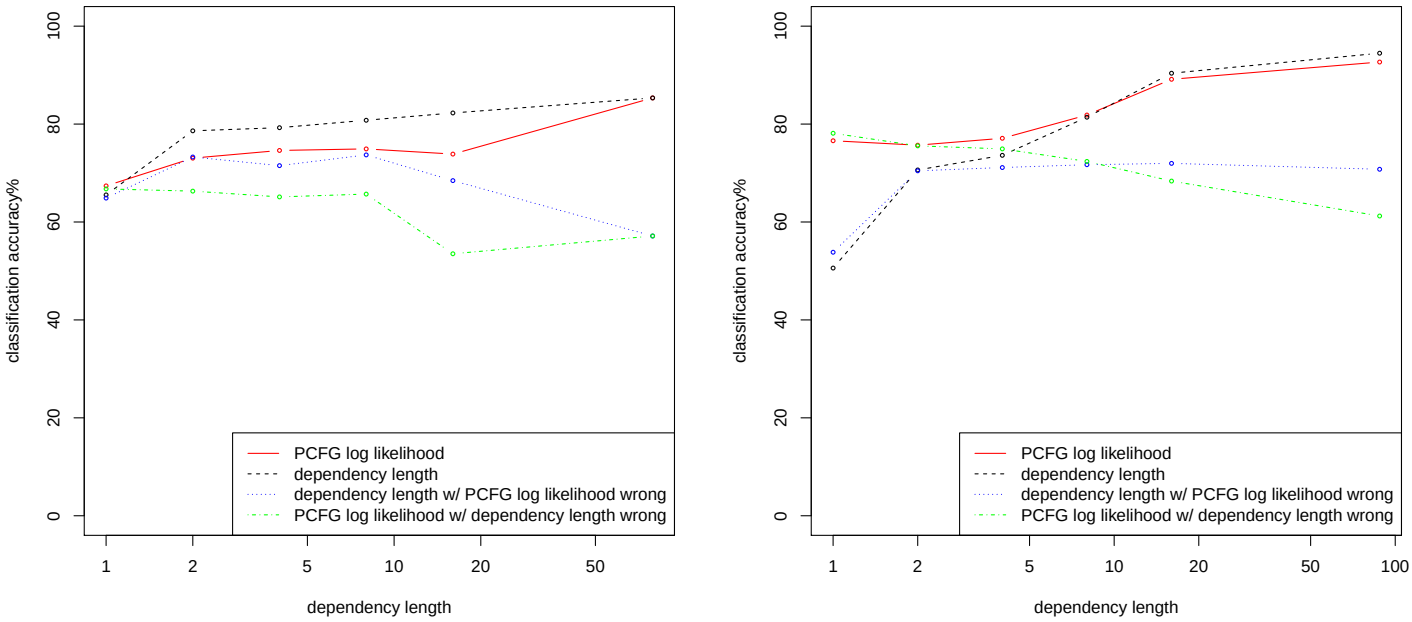


Figure 5: Ablated classification accuracies, with McNemar's χ-square test significance against the previous bar

5.3.1 Accuracy by Dependency Length

Given the distribution of dependency lengths, to avoid data sparsity we divide the syntactic choice pairs into six logarithmically sized bins by the absolute value of the dependency length difference and calculate the prediction accuracy of dependency length and latent-variable PCFG surprisal in each bin, as well as their accuracy on cases where the other predictor



(a) Brown corpus

(b) WSJ corpus

Figure 6: Classification accuracy by binned absolute value of dependency length difference

Deplen range	Accuracy		dependency length acc w/ PCFG log likelihood false	PCFG log likelihood acc w/ dependency length false
	PCFG log likelihood	dependency length		
1 (2121)	68.18	65.54	60.44 (675)	63.47 (731)
2 (1291)	73.66	78.62	67.65 (340)	60.14 (276)
$2 < \text{len} \leq 4$ (1355)	75.50	79.26	66.87 (332)	60.85 (281)
$4 < \text{len} \leq 8$ (1077)	75.77	80.78	70.88 (261)	63.28 (207)
$8 < \text{len} \leq 16$ (643)	75.12	82.27	63.75 (160)	49.12 (114)
$16 < \text{len}$ (191)	85.64	86.14	55.17 (29)	53.57 (28)

(a) Brown corpus

Deplen range	Accuracy		dependency length acc w/ PCFG log likelihood false	PCFG log likelihood acc w/ dependency length false
	PCFG log likelihood	dependency length		
1 (4969)	76.80	50.59	52.90 (1153)	77.88 (2455)
2 (2172)	76.01	70.63	67.18 (521)	72.20 (638)
$2 < \text{len} \leq 4$ (2873)	77.97	73.62	67.30 (633)	72.69 (758)
$4 < \text{len} \leq 8$ (3173)	82.67	81.41	67.82 (550)	70.00 (590)
$8 < \text{len} \leq 16$ (2663)	89.63	90.39	68.12 (276)	65.63 (256)
$16 < \text{len}$ (887)	92.89	94.48	69.84 (63)	61.22 (49)

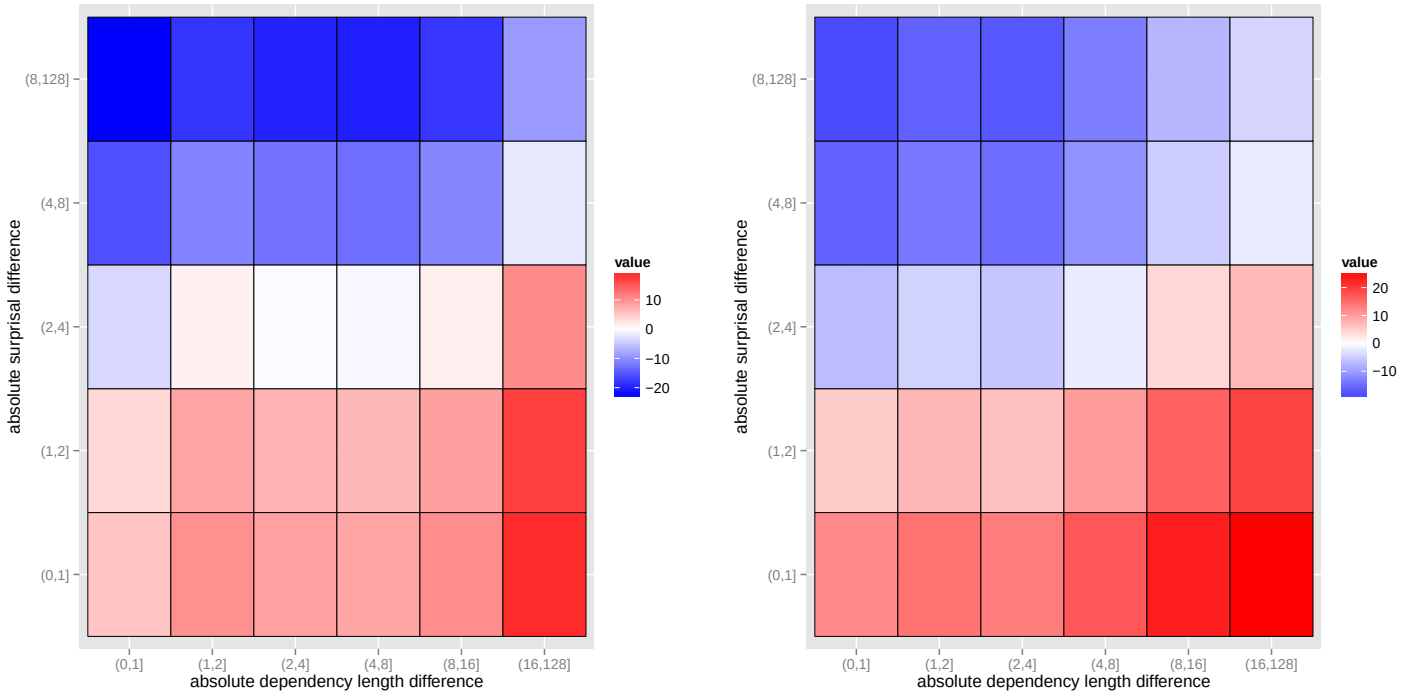
(b) WSJ corpus

Table 7: Classification accuracy by binned absolute value of dependency length difference, with number of data points in parentheses

makes the incorrect prediction (Table 7 and Figure 6). Across the corpora, the prediction accuracy of both measures rises gradually with the increase in dependency length difference. In contrast to Demberg & Keller's results with reading comprehension though, dependency length is generally effective in predicting syntactic choice from the second bin onwards. In particular, in cases where latent-variable PCFG surprisal did not provide the correct prediction, the accuracy of dependency length is well above random chance (50% accuracy) beyond the first bin, except in the final bin with the Brown corpus which has relatively few items. Not surprisingly, in cases where dependency length makes the wrong prediction, latent-variable PCFG surprisal also does well in most bins.

As the table and figure show, dependency length is relatively more effective with the Brown corpus, especially for the smaller sized bins. In Section 5.4, we show that this difference stems in large part from the unequal distribution of constructions across the two corpora. Conversely, latent-variable PCFG surprisal is more effective with the WSJ corpus, as expected given that both parser training and test data are from the same domain in this case.

To better visualize the relative performance and complementarity of dependency length and surprisal, we constructed heatmaps that depict the



(a) Brown corpus

(b) WSJ corpus

Figure 7: Heatmaps depicting difference between classification accuracy of dependency length and latent-variable PCFG surprisal

difference in classification accuracy between the two measures as a function of bins representing the absolute values of both dependency length and surprisal difference (Figure 7). As the figure shows, dependency length is relatively more effective not only with larger differences in dependency length, but also with smaller differences in latent-variable PCFG surprisal.[15]

[15] As one of the reviewers pointed out, Demberg and Keller (2008) results suggest that the effect of dependency length might be non-linear (if memory decay is exponential this is along expected lines). As such, we also tried out dependency length as a quadratic term in the GLM. However, this did not turn out to be a significant predictor of syntactic choice.

5.3.2 Comparison with Demberg and Keller (2008)

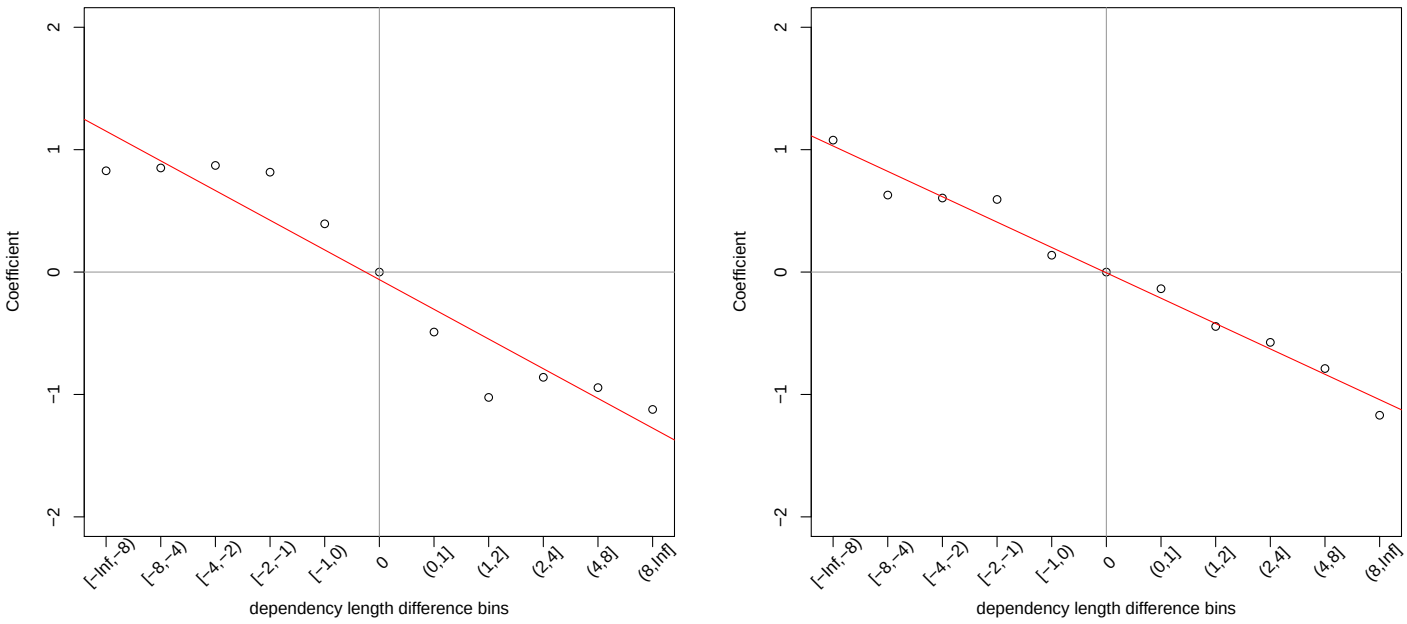
In their work on the Dundee reading time corpus, Demberg and Keller (2008) similarly observe that dependency length is an increasingly effective predictor as dependencies become longer. However, like van Schijndel and Schuler (2013) and van Schijndel et al. (2013), Demberg & Keller report that dependency length has a negative coefficient for integration costs between 0 and 9. Unlike other studies that have found a negative integration cost coefficient though, Demberg & Keller found that for integration costs greater than 9, dependency length induces greater reading times (positive coefficients), as expected. Note that in their work, overall dependency length has negative coefficients because of the preponderance of short dependency length cases in the Dundee corpus: if positive integration cost is only slightly predictive at long distances, the model can shift the entire dependency length regression down to account for it and compensate by shifting up the other lines, such as surprisal, thereby producing a negative integration cost for short and moderate values of dependency length. In any case, the absence of the expected effect except for rather long dependencies in a model that includes frequency-based controls indicates that dependency length is at best a rather weak predictor in the case of sentence comprehension.

In order to facilitate a more direct comparison with Demberg & Keller's results, we also experimented with a regression model using binned dependency length, again using logarithmically sized bins to avoid data sparsity:

$$\begin{aligned} \textit{choice} \sim\ & \textit{PCFG log likelihood} + \textit{lm} + \textit{binned dependency length} \\ & + \textit{weighted embedding depth} + \textit{1-best embedding depth} \end{aligned} \tag{9}$$

The regression coefficients for the bins are plotted against the dependency length differences in Figure 8, along with their line of best fit. The results show a robust, consistent preference for relatively lower dependency length in syntactic choice: lower values of dependency length difference (in this case negative values) have a positive regression coefficient, while higher dependency length difference values consistently get a negative regression coefficient.[16] For the Brown corpus, coefficients at all bins are significant ($p < 0.05$), while for the WSJ corpus, all bins except $[-1, 0)$ and $(0, 1]$ are significant. The binned dependency length model results in only a slight increase in classification accuracy (though not significant) compared to the

[16] In Demberg & Keller's study, dependency length did not contain any negative values, unlike in our case where the regression was performed by calculating the difference in values between the reference and each variant.



(a) Brown corpus

(b) WSJ corpus

Figure 8: Regression coefficients obtained after binning dependency length

unbinned version discussed earlier. Across the range of all dependency length bins, both these models also predict the actual proportions of correct choice in the dataset very closely (see Figure A.4 in Appendix A).

5.4 Construction-Specific Experiments

In psycholinguistics, construction frequencies have been linked to processing difficulty. For example, the relative comprehension ease of subject relative clauses compared to object relative clauses (Gibson 2000) is attributed to the fact that subject relative clauses are more frequent in language compared to object relative clauses (MacDonald 1994; 1999). In this section, we contrast the performance of frequency-based measures against memory-based ones for the task of predicting syntactic choice in various constructions.

Construction	Frequency	Deplen Coefficient	Other Significant Predictors
Dative alternation	538	-6.87, $p = 6.3e-12$	ngram log likelihood
Postverbal adjuncts	5588	-23.46, $p < 2e-16$	PCFG and ngram log likelihoods, weighted embedding depth
Preverbal adjuncts	1656	7.74, $p = 9.7e-15$	PCFG and ngram log likelihoods, weighted embedding depth
Quotations	603	-0.37, $p = 0.71$	PCFG and ngram log likelihoods

(a) Brown corpus

Construction	Frequency	Deplen Coefficient	Other Significant Predictors
Dative alternation	1143	-6.94, $p = 3.9e-12$	PCFG and ngram log likelihoods
Postverbal adjuncts	11966	-27.40, $p < 2e-16$	PCFG and ngram log likelihoods, weighted and 1-best embedding depths
Preverbal adjuncts	3156	11.39, $p < 2e-16$	PCFG and ngram log likelihoods, weighted and 1-best embedding depths
Quotations	4065	-5.70, $p = 1.2e-08$	PCFG and ngram log likelihoods, 1-best embedding depth

(b) WSJ corpus

Table 8: Construction-wise regression

5.4.1 Regression on Constructions

Recent work has investigated the relationship between processing difficulty and frequency (van Schijndel & Schuler 2013) in the framework of Phillips's (2013) grounding hypothesis, namely that high frequency constructions are strategies that languages develop in order to avoid possible downstream processing costs. Presumably, therefore, low frequency constructions would incur a heavier memory load in comparison to their high frequency counterparts. Van Schijndel & Schuler model reading times in written English by incorporating both frequency measures (surprisal and entropy reduction) and memory-based costs (weighted embedding difference and other predictions made by left-corner parsing operations). They showed that memory-based measures are significant predictors of reading time data even when frequency-based measures are considered as controls in the statistical model. We examine the impact of frequency- and memory-based measures on predicting syntactic choices belonging to four distinct constructions (and their subtypes) by running the regression model introduced earlier in Equation 8 on each of the construction types.

The results of regression modelling (Table 8) indicate that at least one

of the memory-constraint measures (dependency length and the left-corner parser measures) is significant for all construction types in both our datasets even in the presence of powerful frequency-based controls (latent-variable PCFG and n-gram surprisal). It is also worth noting that for both datasets, dependency length has a positive regression coefficient for preverbal adjuncts (all other constructions display a negative coefficient for dependency length). This means that for this construction, instead of the tendency towards dependency length minimization, the language has the opposite preference, i.e. increasing dependency length values predict syntactic choice (referred to as non-locality cases henceforth). It is conceivable that there are discourse factors affecting the frontedness of these adjuncts. Temperley (2007) also reports such cases in his corpus study. In the discussion section, we will focus on the efficacy of surprisal and the left-corner memory measures in predicting syntactic choice in non-locality cases.

For the Brown corpus, dependency length (a memory-based measure) does have a significant impact on syntactic choice for all constructions except quotations. But dependency length is a significant predictor of syntactic choice in all WSJ constructions including quotations. Compared to the Brown corpus, WSJ does have a larger number of quotation cases, so this exception may be due to a lack of statistical power. The next subsection in this section compares these two corpora in terms of the number and distributions of these constructions. As we show there, distributional factors are also responsible for the differential classification performance of dependency length across the two datasets.

5.4.2 Classification Accuracy by Construction and Corpus

In this section, we examine the impact of the distribution of constructions across corpora on the classification accuracy of dependency length. We also provide corresponding figures for latent-variable PCFG surprisal for purposes of comparison. For both corpora, PCFG surprisal results in high classification accuracy for all constructions (see Figure 9). In contrast, the performance of dependency length is more mixed. In both datasets, model performance on the dative alternation and postverbal adjuncts is very high.[17]

[17] As one of the reviewers pointed out, in most constructions, the number of words is identical across the reference sentence and the variants. However, in the dative alternation, one of the variants differs with the reference by one word (*to*). We do concede that language models have a bias towards preferring sentences with fewer words. Averaging feature values by dividing by the number of words results in a substantial drop in the classification performance (more than 10%) of all the predictors including language model scores. Hence we report only results obtained from unaveraged version of all measures.

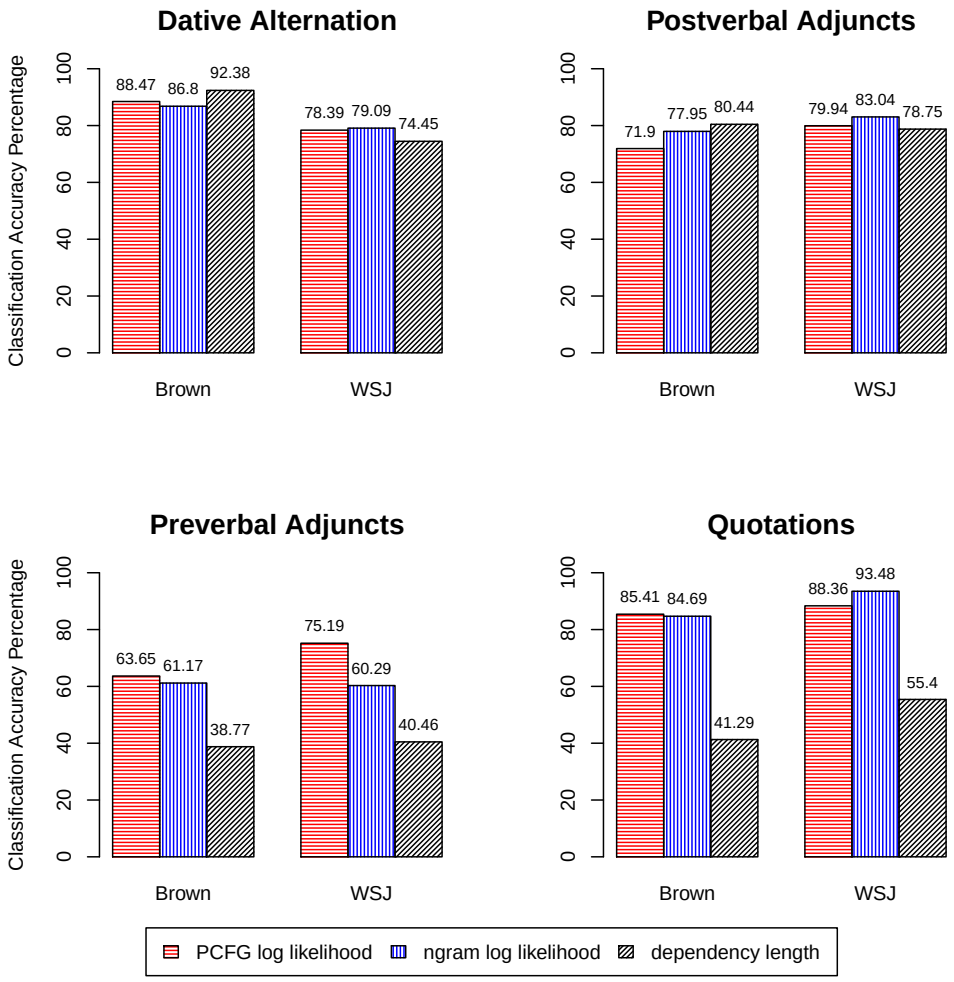


Figure 9: Classification accuracy by construction in (Brown, WSJ) corpora

However, dependency length has classification accuracy much less than random chance (50% accuracy) for preverbal adjuncts and for quotations in the Brown corpus. For both PCFG surprisal as well as dependency length, each construction-specific accuracy level in the Brown corpus is significantly different from the corresponding accuracy value in the WSJ corpus (Bonferroni correction applied for the χ^2 test of correct vs. incorrect number of cases for each accuracy value).

Deplen range #comparisons (Brown,WSJ)	Construction	Brown %cases (acc)	WSJ %cases (acc)
1	Dative alternation	5.18 (78.18)	7.98 (66.75)
2121, 4969	Quotation	8.77 (26.88)	33.38 (24.35)
	Postverbal adjuncts	69.31 (76.05)	46.57 (67.33)
	Preverbal adjuncts	16.74 (38.31)	12.05 (47.92)
	Total	100 (65.54)	100 (50.59)
$1 < \text{len} \leq 4$	Dative alternation	12.51 (97.88)	7.81 (81.47)
2646, 5045	Quotation	0.34 (55.55)	4.10 (53.62)
	Postverbal adjuncts	63.87 (89.17)	65.33 (82.55)
	Preverbal adjuncts	23.28 (41.07)	22.75 (43.21)
	Total	100 (78.95)	100 (72.33)
$4 < \text{len}$	Dative alternation	3.92 (100.00)	3.15 (90.57)
1911, 6723	Quotation	0.26 (80.00)	20.60 (95.67)
	Postverbal adjuncts	70.74 (97.93)	60.86 (97.80)
	Preverbal adjuncts	25.06 (33.19)	15.38 (29.89)
	Total	100 (81.74)	100 (86.69)

Table 9: Distribution of preverbal adjunct and quotation cases and dependency length accuracy across 3 dependency length difference bins

We seek to explain this differential performance of dependency length across the two corpora by comparing the distribution of constructions in each. The overall distribution of constructions across the two corpora are significantly different (4x2 contingency table; $\chi^2 = 727.15, df = 3, p < 2.2e-16$). To investigate further, we performed a more fine-grained analysis by examining the performance of dependency length along various dependency length bins. Three logarithmic bins of absolute dependency length ranges were created and classification accuracy of dependency length was calculated in each of these bins. Dependency length performance is lower for the smaller dependency length difference bins in the WSJ corpus compared to the Brown corpus. Table 9 illustrates this along with the distribution of constructions inside these bins of interest.

As mentioned earlier, postverbal adjuncts and dative alternations are constructions exhibiting a larger number of locality cases compared to quotations and preverbal adjuncts. In particular, preverbal adjuncts involve frame adverbials and fronting of adjuncts. Here, corpus sentences themselves have the long-short constituent order as Temperley (2007) discusses. For the length-1 bin, compared to the Brown dataset, the WSJ corpus has fewer postverbal adjunct cases and there also the classification performance of dependency length is lower. For this bin, the overall construction distributions across the 2 corpora are significantly different (4x2 contingency table; $\chi^2 = 530.75, df = 3, p < 2.2e-16$) and individual accuracies except in dative alternation cases are also significantly different (Bonferroni correc-

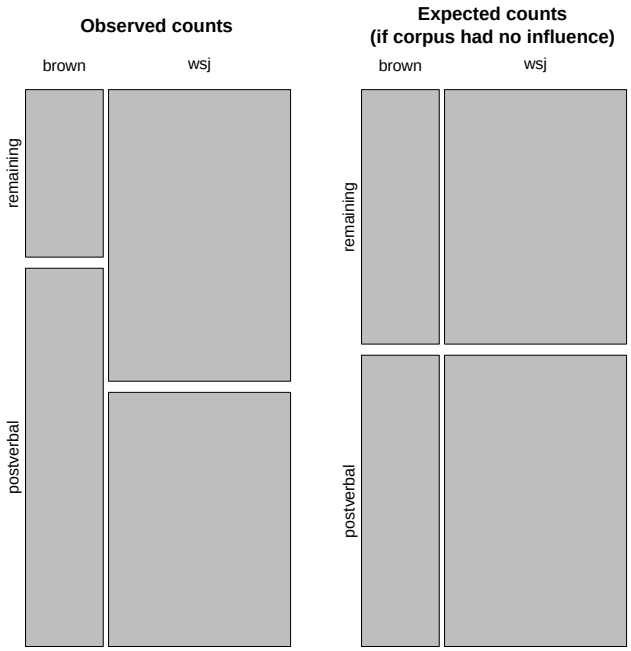


Figure 10: Distribution of postverbal adjuncts vs. all remaining constituents in bin where dependency length difference is 1

tion applied for the χ^2 test of correct vs. incorrect number of cases for each accuracy value).

Investigating this further, we divided the length-1 bin into 2 classes: postverbal adjuncts and all the remaining constructions. Here also, overall both groups are significantly different (2x2 contingency table; $\chi^2 = 307.91, df = 1, p < 2.2e-16$). The observed number of postverbal adjunct cases in the length-1 bin of WSJ is less than the expected number based on the Brown corpus, as Figure 10 shows. This is a plausible explanation for the 14% overall accuracy difference between both corpora. For the middle bin, the accuracy difference narrows down to 6.5%. All constructions are approximately equally distributed in both corpora in this bin and the overall accuracy difference can only be accounted by the slightly lower performance of dependency length in all except one WSJ construction type. However, in the third bin (length > 4), WSJ contains more quotation cases (accuracy is 95% for this construction) and fewer preverbal adjuncts compared to the Brown corpus. This is a plausible explanation for why in this bin the WSJ

corpus exhibits almost 5% greater classification accuracy over the Brown corpus. On the basis of this analysis, we conclude that the differences in the distributions of constructions across corpora is an important factor affecting the performance of dependency length in syntactic choice.

6 Discussion of Dependency Locality

Experiments in the previous section showed dependency length is only a strong positive predictor of syntactic choice on rightward dependencies, with effectiveness increasing with length. This section explores possible reasons for this with the help of linguistic examples.

6.1 Efficacy of Dependency Length

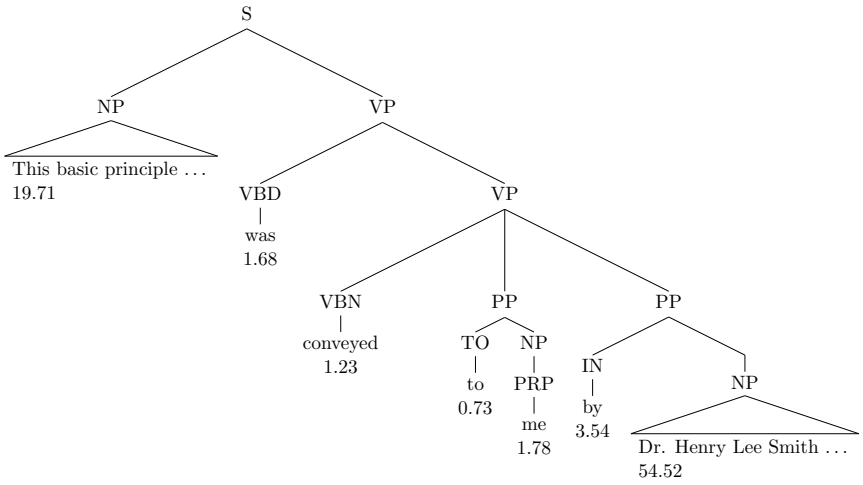
Dependency length is effective when the difference between the dependency lengths of the reference sentence and the variant is at least moderate, and it is most effective when the difference is large (more than 8 discourse referents), as shown in Section 5 (Table 7). This lends further support to the conjecture expressed by Levy (2008) that comprehension difficulty arising from integration of long distance heads is intrinsically different from difficulty arising from predictions of next words given lexical or syntactic context, which surprisal quantifies.[18] In our data, this pattern is most pronounced in the case of verbs involving multiple prepositional phrases as in the example below. Compared to the variant (b), the reference sentence (a) has a much lower value for dependency length (53 vs. 65) but a slightly higher value for latent-variable PCFG surprisal (263.99 vs. 263.57).

(15) a. This basic principle, the first in a richly knotted bundle, was conveyed to me by Dr. Henry Lee Smith, Jr., at the University of Buffalo, where he heads the world's first department of anthropology and linguistics. (CF01.11.0)

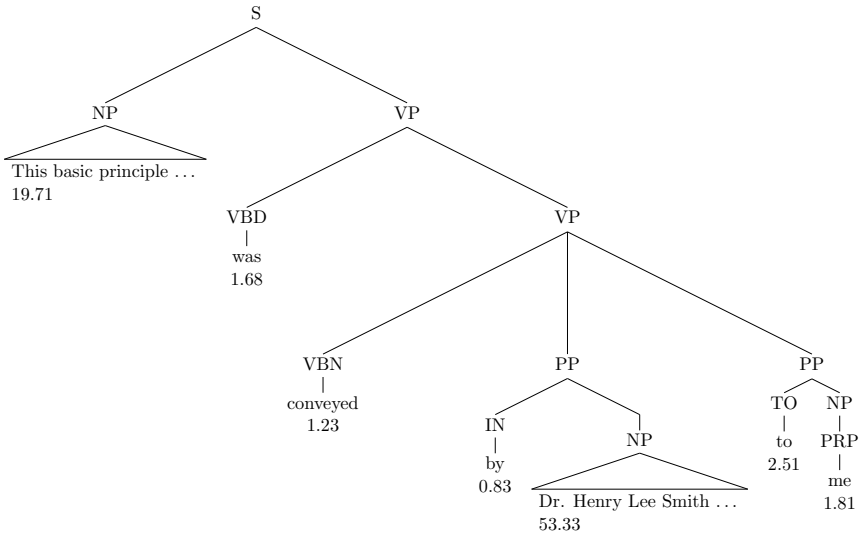
b. This basic principle, the first in a richly knotted bundle, was conveyed by Dr. Henry Lee Smith, Jr., at the University of Buffalo, where he heads the world's first department of anthropology and linguistics to me.

Thus the parser displays a subtle preference for the variant. In total in the WSJ training data for the parser, there are 119 *by-to* orders as opposed to

[18] Note, however, that in our case the dependencies are always syntactically local, even if linearly distant.



(a) Reference sentence



(b) Variant sentence

Figure 11: Parse trees for reference sentence and variant, with incremental surprisal for each word/constituent indicated underneath

95 *to-by* orders. Specifically, for passives (detected by *by*-phrases involving NPs), there are 70 instances of the *by-to* pattern, while the *to-by* pattern is found only in 34 instances. This preference is also illustrated in Figure 11 where the parses have the same surprisal until the verb *conveyed*, but start differing at the preposition.[19] In contrast, dependency length straightforwardly predicts the short-long constituent order in the reference sentence as opposed to the long-short pattern in the variant. The weakness of parser probabilities in this case is not surprising, since the probability of preposition attachment to a verb phrase is roughly equivalent in both cases.

6.2 Divergences from Dependency Length Minimization

From our results it is clear that English demonstrates a clear preference towards minimization of dependency length in constituent ordering decisions. However, there are situations where this tendency is overridden. In this section, we discuss in detail two such situations where the tendency for dependency length minimization is not effective in distinguishing between the reference sentence and the generated variant:

1. *Zero dependency length difference cases*: Here both reference and variant sentences have the same dependency length.

2. *Non-locality cases*: The literature discusses several cases, notably adverb placement and preverbal adjuncts, where divergences from orders predicted by dependency length minimization occur, i.e. when constituent orders with greater dependency length are preferred over a lower dependency length variant (Gildea & Temperley 2007, Temperley 2007).

These cases are attested when certain other factors override dependency length minimization. Hawkins (2014) characterized the interplay between different factors influencing a linguistic choice as belonging to 3 types: 1. Pattern of Preference 2. Pattern of cooperation and 3. Pattern of competition.

Each factor has an individual strength in a given direction (reflected in regression model coefficients and sign in this work), while factors also reinforce each other in many cases. At the same time, competition between locality and other factors result in word order divergences which do

[19] Here we used an incremental parser discussed in (van Schijndel et al. 2013) which also uses the same split-state latent-variable grammar. This parser emits per-word surprisal as opposed to a global likelihood.

not respect locality constraints. Non-locality cases arise as a by-product of competing factors like animates-first, lexical-semantic dependencies, given precedes new information status considerations and topic prominence discussed in Section 3 as well as previous work Hawkins (2004; 2014). In the ensuing discussion, we focus on the impact on these cases of our memory-based measures other than dependency length from a quantitative as well as qualitative perspective. It is an untested hypothesis that the latent-variable PCFG grammar we used to estimate surprisal models all these competing motivations. But the following section outlines plausible reasons why PCFG surprisal estimated by the latent-variable parser actually might be potentially effective in modeling constituent order.

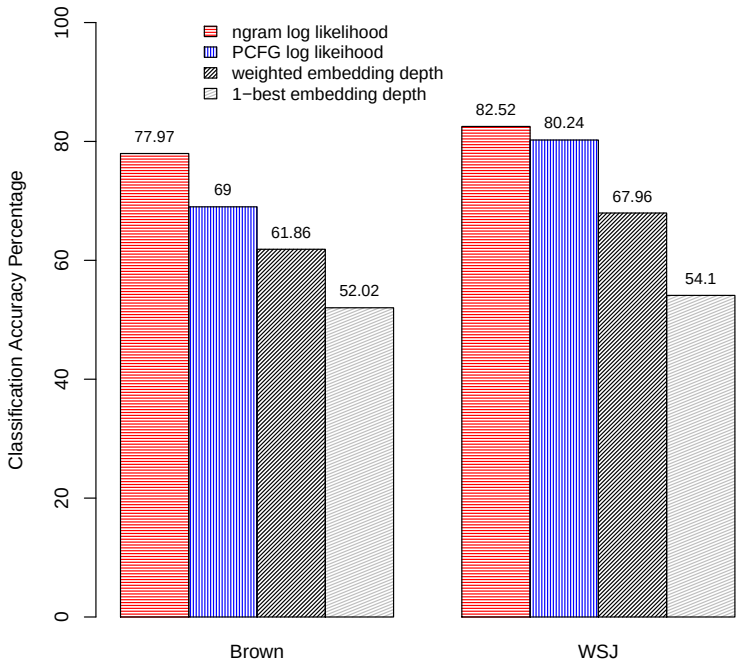
6.2.1 Zero Dependency Length Difference Cases

For equal dependency length cases, the classification accuracy of our other memory-based measures reveals that weighted embedding depth is a close competitor to latent-variable PCFG surprisal followed by embedding depth measure (Figure 12a). However, the left-corner measures do not significantly increase classification accuracy of predicting the correct choice over n-gram and PCFG surprisal in these cases of equal dependency length (Figure 12b). The success of surprisal is illustrated using the following example from the Brown corpus, where both the reference sentence and the variant have total dependency length of 24:

(16) a. He turned over impatiently and pulled the sheet over his head against the treacherous encroachment of the dawn. (cr04.36.0)
b. He turned impatiently over and pulled the sheet over his head against the treacherous encroachment of the dawn.

Here the reference sentence has a lower latent-variable PCFG surprisal of 124.26 compared to 126.71 for the variant, thus sentence likelihood is a better predictor of the reference sentence.

The success of surprisal (in terms of classification performance) points to the need for more detailed analyses of the latent-variable PCFG grammar we used to estimate syntactic surprisal. We now provide some preliminary evidence that our current surprisal measure is effective in modeling some of the factors influencing constituent ordering discussed before in Section 3. In our work, the latent-variable grammar creates many splits for nominal categories as evinced by the most frequent words in several subcategories. Pronouns and nouns have many subcategories, and definiteness information of NPs when lexically marked is also signified by fine-grained determiner



(a) Individual classification accuracy

Model	Brown	WSJ
PCFG+ngram log likelihoods	79.14	85.28
PCFG+ngram log likelihoods+ weighted+1-best embedding depths	79.20; $p = 1$	85.42; $p = 0.6148$

(b) Collective classification accuracy

Figure 12: Individual and collective classification accuracy in zero dependency length cases for Brown (1707 data points) and WSJ (3593 data points) corpora

categories for *the*, *a*, *this* and *some* (see Table 1 of Petrov et al. 2006). Thus lexically marked discourse status is taken into account to a great extent using fine-grained categories. PCFG surprisal models animacy to a certain extent since the fine-grained categories inferred by the latent-variable grammar encode several distinctions based on proper nouns and company names (Petrov et al. 2006). The latent-variable grammar used to estimate PCFG surprisal can potentially model lexical bias in our study. Verb phrases receive subcat-

egories corresponding to infinitive VPs, passive VPs, intransitive VPs and those with sentential and NP/PP complements. These phrasal rules also interact with lexical splits, as the two most frequent rules involving intransitive verbs in our trial were VP-14 → VBD-13 and VP-15 → VBD-12, where VP-14 was associated with a main clause while VP-15 was associated with subordinate clause VPs. In our work, the latent-variable grammar encodes the distinction between verbal arguments and adjuncts to a substantial extent. For example, Petrov et al. (2006) mention the fact that the iterative training procedure involved in estimating the latent-variable grammar ignores some classes of adverbs to learn more generic rules like VP-2 → VP-2 ADVP-6, where the rule VP-2 is not changed in the result due to the addition of ADVP-6. More detailed quantitative investigations are imperative in order to concretely establish the contribution of surprisal in modeling each of the factors mentioned above.

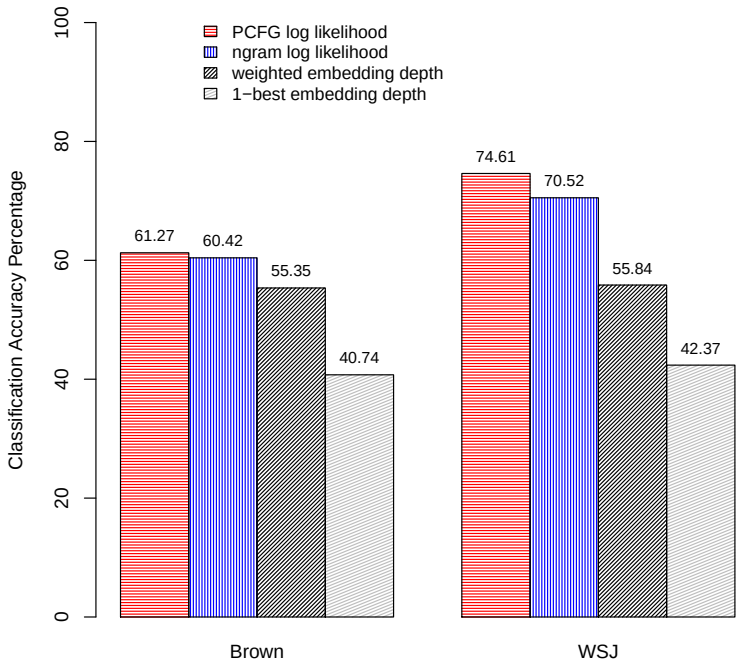
6.2.2 Non-locality Cases

In the non-locality cases in both corpora, latent-variable PCFG surprisal performs best individually, followed by n-gram surprisal and then the other memory measures (Figure 13a). For both corpora, a model containing all the left-corner parsing measures does induce a significant improvement in classification accuracy over a model containing just the two frequency-based measures (Figure 13b).

Next we turn our attention to two constructions involving non-locality cases that have also been discussed in the literature: (i) facts from adverb placement (Gildea & Temperley 2007), and (ii) data from sentence initial premodifying adjuncts (Temperley 2007).

Adverb Placement Gildea and Temperley (2007) suggest that adverb placement might involve cases which go against dependency length minimization. Pursuing this suggestion, we examined 295 legitimate long-short postverbal constituent orders from Section 00 of the Penn Treebank. Table 10 shows the distribution of second constituent function tags in these sequences. The proportions indicate that there is a predominant tendency for the shorter constituent to express temporal information. Both PCFG and n-gram surprisal are effective in such examples, as illustrated below:

(17) a. When the Half Moon put in at Dartmouth, England, in the fall of 1609, word of Hudson's findings leaked out, and English interest in him revived. (CF16.25.0)



(a) Individual classification accuracy

Model	Brown	WSJ
PCFG+ngram log likelihoods	64.45	76.52
PCFG+ngram log likelihoods+ weighted+1-best embedding depths	66.34; $p = 0.00037$	77.33; $p = 6.6e-05$

(b) Collective classification accuracy

Figure 13: Individual and collective classification accuracy in non-locality cases for Brown (1637 data points) and WSJ (4746 data points) corpora

b. When the Half Moon put in in the fall of 1609, at Dartmouth, England, word of Hudson's findings leaked out, and English interest in him revived.

Here the reference sentence has a dependency length of 47 while the variant has lower dependency length of 46 discourse referents. But the reference has a higher parser log likelihood and language model score (hence lower PCFG and n-gram surprisal values) compared to the variant. In contrast, memory-

Function Tag	%Short 2^{nd} Constituents
TMP	42.4
CLR	12.2
LOC	11.2
PRP	4.4
ADV	4.4
DIR	4.4
MNR	3.05

Table 10: Distribution of the second function tag in 295 long-short sequences in PTB Sect. 00

based measures are not effective in predicting the reference sentence. The efficacy of frequency-based measures in these cases is due to their bias towards more frequent lexical and syntactic patterns in the data on which these measures were estimated. In the above examples, the reference sentence contains the phrasal verb *put in* followed by two constituents headed by *at* and *in* (the second constituent expressing temporal information as per the trend discussed above). In contrast, the variant has the competing order of postverbal constituents having heads *in* and *at*. The success of surprisal can be attributed to the fact that the WSJ sections on which the parser was trained contain 117 instances of the *at ... in* sequence of postverbal constituent heads, as opposed to merely 51 instances of the *in ... at* sequence seen in the variant. Similarly, the language model disprefers the *in in* bigram sequence found in the variant compared to the more frequent *in at* bigram in the reference sentence.

Premodifying Adjuncts A closer look at Temperley's (2007) original corpus study also revealed counter-examples to dependency length minimization. A case in point is the class of examples involving premodifying adjunct sequences that precede both the subject and the verb. As mentioned in Section 5, dependency length displays a positive coefficient for preverbal adjuncts (as opposed to a negative coefficient for all other constructions).

In the case of premodifying adjuncts, assuming that the parent head is the main verb of the sentence, a long-short sequence would minimize overall dependency length. However, as Temperley (2007) reports, in 613 examples found in the PTB, the average length of the first adjunct is 3.15 words while the second adjunct is 3.48 words long, thus reflecting a short-long pattern.

In the Brown Corpus, the length difference is more pronounced (2.44 words for first adjuncts vs. 4.22 for second adjuncts on average). The following examples illustrate this (David Temperley p.c.):

(18) a. [In 1976], [as a film student at the Purchase campus of the State University of New York], Mr. Lane, shot "A Place in Time", a 36-minute black-and-white film ...(WSJ0039.4)

b. [As a film student at the Purchase campus of the State University of New York] [in 1976], Mr. Lane, shot "A Place in Time", a 36-minute black-and-white film ...

Informal native speaker judgements indicate that the variant sentence (18-b) above, which minimizes dependency length, is less preferred compared to the original corpus sentence (18-a). The frequency-based measures of PCFG and language model surprisal correctly prefer the reference sentence. Thus, in the sentence initial position, speakers might be overriding the tendency to minimize dependency length as a consequence of other considerations. However it is worth noting that surprisal is not effective in all such cases as exemplified below:

(19) a. Then, as an additional precaution, the car dealership took the judge's photograph as he stood next to his new car with sales papers in hand – proof that he had received the loan documents. (WSJ0267.78.0)

b. As an additional precaution, then, the car dealership took the judge's photograph as he stood next to his new car with sales papers in hand – proof that he had received the loan documents.

Here, the reference sentence has dependency length 59 and parser surprisal 217.76 while the variant has corresponding values 58 and 217.73. Thus both these measures prefer the variant. The reason for this might be the fact that the WSJ sections used for parser training have 444 instances of preverbal adjunct sequences headed by prepositions and adverbs (*IN-RB* tags) as opposed to only 190 instances of the opposite *RB-IN* sequence. Thus it is possible that the frequency bias of surprisal can be detrimental to predicting the correct choice. In contrast, for many of these cases where surprisal is not effective, the left-corner memory-based measures of embedding depth prefer the reference sentence over the variant.

A competing explanation for non-locality cases would be the "short-first" principle proposed by Arnold et al. (2000). This principle states that constraints in the production system result in a preference for realizing short

constituents first. So at all choice points, short constituents are considered to be easier to produce in comparison to longer (and hence more difficult) constituents, which speakers postpone until later points in the production stream. Thus this formulation predicts an overall preference for short-long constituent orders across the board (both preverbally as well as postverbally). As a consequence, preverbal non-locality cases are accounted for directly by the "short-first" principle, as will all other DLT predictions for English (including postverbal short-long constituent orders). However, as Temperley (2007) argues, this formulation fails to account for the predominant pattern of long-short preverbal constituent orders observed in head-final languages like Japanese and Korean (Choi 2007, Yamashita & Chang 2001). In contrast, these patterns associated with head-final languages fall out directly from DLT predictions, making DLT an attractive account of processing with cross-linguistic plausibility. In addition to Arnold et al.'s "short-first" principle, there is the general framing or topic-first preference, which is seen in conventionalized form in "topic-prominent" languages such as Chinese and Japanese (Hawkins 2004; 2014). These works discuss a variety of cases involving the topic (expressed as the adjunct) and the constituent expressing the main predicate. As Hawkins (2004) states, the "frame-setting" topics contribute to the enrichment of the predicate. These enrichments can be in terms of expressing spatial, temporal and causation information via the topic (Hawkins 2014). Thus ordering the topic before the main predicate helps reduce the possibility of semantic misassignments. It remains for future work to establish whether theories of working memory like ACT-R can provide unified explanations for both locality and non-locality cases in production along the lines of previous results for language comprehension (Vasishth & Lewis 2006).

Finally, an alternative that merits further research is that discourse considerations predominate in choosing initial sentence elements. In NLG with German, Filippova and Strube (2007) find it useful to separately choose the initial constituent of the sentence prior to all other constituent ordering choices. In the examples discussed above, (18-a) begins with a FRAME ADVERBIAL (see Maienborn 2001), an adverbial that serves to establish a frame (or set the scene) for the ensuing event description. With such adverbials, it seems plausible that their discourse function would override the concerns of the memory-based measures investigated here. Meanwhile, example (19-a) begins with a DISCOURSE ADVERBIAL (Webber 2004; 2006, Webber, Stone, Joshi, & Knott 2003), a connective which involves an anaphoric dependency to an element in the prior discourse context in addition to the syntactically-mediated dependency to the main verb. Since DLT (as formulated here)

does not take into account anaphoric connections, it would appear fruitful to investigate in future work memory-based measures that do include such anaphoric dependencies.

As of now discourse considerations do not feature in any of our predictors. In language, discourse connectives are function words which perform a variety of functions that help the reader/listener comprehend the message conveyed by the speaker/writer effectively. Discourse connectives establish coherence links between textual spans as well as facilitate inferencing during the interpretation process. Future work can investigate the possibility of integrating a computational model of discourse relations where surprisal is calculated over discourse relations from the Penn Discourse Treebank (PDTB) resource (Prasad et al. 2008) which classifies discourse relations into four broad types: Temporal, Contingency, Comparison and Expansion. Preliminary evidence for the information theoretic basis of discourse marker identity and mentioning arises from corpus studies conducted by Vera Demberg and colleagues. According to the UID hypothesis, when the relationship between two textual units is not along expected lines (not easily predictable) discourse connectives are overtly mentioned so that the overall information density is uniformly distributed. Conversely, when the relations between textual units are predictable, the connective is implicit (not overtly mentioned). Based on the above insight, Asr and Demberg (2012) test and confirm the following hypothesis about continuity and causality markers in text. Continuity and causality discourse markers are implicit more frequently than other discourse markers. They identified certain TEMPORAL and all COMPARISON markers as encoding discontinuity while most EXPANSION markers mark continuity. CONTINGENCY markers are not related to continuity and denote causality instead. They quantified the implicitness of a relation as the ratio between the number of implicit relations and the total number of relations in the PDTB corpus. So a computational model of discourse relations can predict the discourse connective given all these different factors. One possibility is to use a maximum entropy classification model to predict connectives and estimate information density as the classification probability based on contextual factors like measure of information gain, lexical cue strength of preceding words, distance between discourse connective and other lexical cues and syntactic factors like construction type and parallelism.

7 General Discussion

Findings like ours of memory-based influences on production could potentially contribute towards an integrated theory of comprehension as outlined by Pickering and Garrod (2013). They argue against the almost complete separation between theories of language comprehension and production that currently exists in psycholinguistics. Instead, they argue that language production and comprehension occur in interleaved fashion during real-life language use. Pickering and Garrod (2013) also present evidence from behavioural and neural studies that both production and comprehension systems make predictions by taking inputs from each other. Hence information-theoretic measures like surprisal can facilitate quantitative modelling of linguistic interactions in a theoretical framework integrating mechanisms of both production and comprehension.

The hypothesis of audience design proposes that speakers tend to adjust their speech to suit the needs of listeners (Bell 1984), so audience design would predict that there is a tendency to avoid temporary syntactic ambiguities while producing language. In light of our results and those of Arnold (2011), however, it seems unlikely that the language production system is actively seeking to ease comprehension. As Jaeger and Buz (in press) state, communicative ease need not be due to altruism from the end of the speaker whereby they are indulging in audience design to facilitate communication for the listener. Speakers might have their own communicative goals or according to availability-based production accounts they might be realizing the most readily available constituents. Conceivably, cognitive accessibilty might also be inducing realization of the most accessible elements. Although there is evidence of self-monitoring at the phonetic level (W. J. M. Levelt 1989), studies have failed to yield consistent evidence for ambiguity avoidance as a strategy in language production (Arnold 2011, Arnold et al. 2004, Roland, Elman, & Ferreira 2006, Temperley 2003). If our findings of memory-based effects hold up in spoken language, they may best be interpreted as arising from the production process rather than attempts to facilitate communication.

The Production-Distribution-Comprehension (PDC) account by MacDonald (2013) proposes that word order choices are influenced largely by computational constraints of language production like memory retrieval and motor planning. MacDonald discusses the following factors related to production ease: 1. *Easy First* 2. *Plan Reuse* and 3. *Reduce Interference*. The first factor *Easy First* encodes the idea that more accessible elements are realized early or in relatively more prominent parts of the sentence and

this is the source of word order flexibility. The second factor, *Plan Reuse*, in contrast, is conceived as the source of word order rigidity whereby grammatical constraints of the language license certain word orders while blocking certain others. In addition, certain structures are produced since they have been recently uttered in the discourse (structural persistence or syntactic priming). The third factor *Reduce Interference* refers to the tendency of producers to realize words and structures so as to minimize interference with other elements in the utterance plan. Thus some items are inhibited and some others are activated and subsequently produced. Actual forms and structures which are a result of the production process are a product of the interplay between these three factors and cross-linguistic variation is caused due to the relative degree to which these three factors operate in a given language. These choices when repeated over many structures and individuals mould linguistic forms and their changes. However, although models of working memory have been used to explain sentence comprehension phenomena (Gibson 2000, Lewis et al. 2006, Schuler 2014, van Schijndel et al. 2013, Vasishth & Lewis 2006), explanations based on working memory have only recently been used to explain the mechanisms of language production (Martin & Slevc 2014, Reitter et al. 2011, Slevc 2011). For example, in the context of dative alternation choices, Slevc (2011) shows how speakers exploit the flexibility offered by the grammar to choose more accessible syntactic structures which reduce the potential for interference in memory, and Reitter et al. (2011) show how ACT-R can account for syntactic priming in language production. It remains to be established with studies from more languages whether working memory mechanisms like interference and retrieval attested in comprehension processes are indeed germane for syntactic choice in language production as well.

According to PDC assumptions, language perception involves learning the distributional patterns in the production data and using this experience to facilitate comprehension routines. MacDonald thus explains the comprehension ease and difficulty associated with animacy and verb type in relative clauses by linking it to the frequencies of producing these structures (both spontaneous production as well as corpus data). As Levy and Gibson (2013) discuss, surprisal theory is very much synergistic with the PDC approach as it models distributional regularities in production data. PCFG surprisal has the potential to quantify the impact of various factors affecting constituent ordering we discussed in Section 3 as well as real-world frequencies and expressive biases. So we do believe that surprisal is important for accounts of linear ordering and is potentially compatible with explanations of *production ease* as well as *communicative accounts* (Jaeger & Buz in press). Theories of

language production need to account for constituent order patterns in a wide variety of languages. However, a lot of the preferences visible in production data, such as mirror-image weight effects across different (VO and OV) languages, are not actually predicted by current production models (Hawkins 2014). Current accessibility and availability accounts of language production do not predict the long-before-short order in SOV languages (Jaeger & Norcliffe 2009). Even those that advocate strong alignments between production and comprehension do not incorporate mechanisms for showing in any syntactic detail how speakers formulate their syntactic trees so that both short-before-long and long-before-short orders can be advantageous for them as well as for the hearer in different types of languages, and why production data end up looking like what a parser would prefer within a comprehension model.[20] The efficacy of surprisal across various languages with differing degrees of word order freedom needs to be investigated more thoroughly.

The results of our classification experiments quantify the individual and collective merit of several comprehension factors. They can potentially contribute towards cognitively grounded theories about why writers (or speakers, if extended to spoken data) choose a particular sentence while eliminating several other plausible variants. The success of surprisal in modelling syntactic choice data has implications for probabilistic theories of language production. Aylett and Turk (2004) demonstrated that in human language production, predictability of words is related to their durations and articulatory detail. This finding is also compatible with connectionist models of language production (Chang, Dell, & Bock 2006). More recently, probabilistic information has been incorporated into accounts of optional word mention (optional complementizer, contractions and optional case marking). The UNIFORM INFORMATION DENSITY (UID) hypothesis (Jaeger 2010) states that speakers tend to avoid steep peaks or troughs in information density by inserting or avoiding optional *that*-complementizers in English. Though Jaeger's work deals with reduction choices, which are orthogonal to the ordering choices we examine in this work, Jaeger suggests that it might be worthwhile to investigate whether there is a tendency to make information density uniform at all choice points in language production. It might be relevant to test whether the tendency to minimize spikes in surprisal across words or constituents (depending on incrementality assumptions in production) is independently driving linear ordering. In the case of English syntactic choice phenomena, there is also some preliminary evidence that uniform information density (quantified by surprisal differences at successive words)

[20] We are indebted to one of the reviewers for this idea.

is a better predictor of human sentence ratings (Collins 2012). We leave explorations of uniform information density and syntactic choice for future inquiries.

The complementary nature of surprisal and dependency length predictions for both sentence comprehension and syntactic choice in written text have implications for theories of language cognition. Further inquiries can explore the degree and nature of overlap between mechanisms of language comprehension and production, thus contributing to integrated theories. In terms of cognitive modelling, Demberg, Keller, and Koller (2013) emphasize the importance of formulating an integrated measure which combines the predictions made by both these measures. They formulate the Prediction Theory where comprehension costs are calculated by summing syntactic surprisal (cost of updating syntactic structure) and verification cost (cost of integrating predicted structure). The verification cost component is inspired from DLT integration costs and is calculated using an equation having an exponential term which models the extent to which predictions have decayed in memory at the time of verification. Subsequently they show that prediction theory models reading times in the Dundee corpus much better than previously reported surprisal measures reported by Roark et al. (2009). Future inquiries can explore computational models to examine whether the two factors stem from one common underlying preference: to keep linguistic elements that are predictive of each other temporally close in the speech stream.

Finally, the claim that the processing complexity of a construction is influenced by its frequency prompts the question as to why language as a system contains some constructions which are less frequent than others given the same semantics. One explanation which has been proposed in the literature is that some constructions are less frequent because they are more difficult or require more memory to produce (Culicover 2014). This claim has some empirical support from a recent experimental study by Scontras, Badecker, Shank, Lim, and Fedorenko (2015). Using two elicited-production experiments, they show that object-extracted structures (relative clauses and wh-questions containing non-local dependencies) take longer to begin and produce compared to their subject-extracted counterparts (containing only local dependencies). They also report that object-extracted structures induce more disfluencies in comparison to subject extractions. As Culicover (2014) states, there may be a loop connecting production complexity from the speaker's perspective to frequency, and in turn linking frequency to comprehension complexity for the hearer. Thus, it might be fruitful to extend our use of frequency-based predictors from cross construction written

data to manipulated constructions with equivalent meanings in speech data and examine how frequency- and memory-based measures of comprehension correlate with production difficulty (measured by disfluencies and speech repairs). Given the fact that frequency effects show a distinct bias towards patterns common in prior experience, it would be insightful to quantify the role of memory-based measures in offsetting this disadvantage.

8 Conclusions

In this paper, we have shown that dependency length is a significant factor in predicting syntactic choice in written English even when surprisal and other cognitively grounded control variables are present in the regression model. We also report that for syntactic choice phenomena, dependency length and surprisal are only moderately correlated. Thus these measures make complementary predictions and model different parts of the data, with the efficacy of dependency length increasing as head-dependent distances increase. Our results showing the complementary nature of dependency length and surprisal for syntactic choice echo Demberg & Keller's (2008) results for sentence comprehension. However, while attempts to observe the predicted influence of dependency locality on sentence comprehension have met with mixed results (Demberg & Keller 2008, Shain et al. 2016, van Schijndel et al. 2013, van Schijndel & Schuler 2013), the present study provides robust evidence that dependency length is a significant influence on the choice between multiple syntactic alternatives in written English, not only for relatively long dependencies but also those of moderate length. We have also investigated cases where dependency locality systematically fails to make correct predictions, and have shown that some constituent orders that diverge from the general preference for dependency length minimization can be accounted for by the embedding depth measures of comprehension discussed by Wu et al. (2010).

In future inquiries, it will be fruitful to extend this study to spoken language production by using transcribed speech corpora as well as behavioural experiments, enabling us to determine whether the measures considered in this study are also valid for a theory of language production. Previous authors have stated that evidence for many of the pressures observed in spoken language production can also be observed in writing (Jaeger 2011), and Gildea and Temperley (2007) report that both written as well as transcribed speech show very similar dependency minimization patterns. Although transcribed speech data is noisy due to pauses, interjections and speech repairs,

using incremental parsers developed to parse speech data (Miller & Schuler 2008) it should nevertheless be feasible to extend our work to examine the contribution of working memory in actual mechanisms of production using spoken language corpora. It will also be fruitful to investigate the role of more general-purpose theories of working memory like ACT-R, which have been proved effective in language comprehension, on the actual mechanisms of language production. Finally, another promising line of inquiry is investigating the role of discourse context in fronting decisions that go against dependency locality, given that discourse considerations appear to often predominate in such decisions over the memory-based measures pursued here.

Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship [Grant No. DGE-1343012] awarded to the second author. The first author would like to acknowledge assistance from Research Grant for New Faculty Scheme [Grant No. MI01195] by IIT Delhi. This work was also supported by an allocation of computing time from the Ohio Supercomputer Center. We would also like to thank the editor, three reviewers, Vera Demberg and Florian Jaeger for useful commentary and feedback.

References

Anttila, A., Adams, M., & Speriosu, M. (2010). The role of prosody in the english dative alternation. *Language and Cognitive Processes*, *25*(7-9), 946-981. Retrieved from `http://dx.doi.org/10.1080/01690960903525481` doi: 10.1080/01690960903525481

Arnold, J. E. (2008, June). Reference production: Production-internal and addressee-oriented processes. *Language and Cognitive Processes*, *23*(4), 495–527. Retrieved from `http://www.tandfonline.com/doi/abs/10.1080/01690960801920099` doi: 10.1080/01690960801920099

Arnold, J. E. (2011). Ordering choices in production: For the speaker or for the listener? In E. M. Bender & J. E. Arnold (Eds.), *Language from a cognitive perspective: Grammar, usage, and processing* (pp. 199–222). CSLI Publishers.

Arnold, J. E., Wasow, T., Asudeh, A., & Alrenga, P. (2004). Avoiding attachment ambiguities: The role of constituent ordering. *Journal of Memory and Language*, *51*.

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

Asr, F., & Demberg, V. (2012, December). Implicitness of discourse relations. In *Proceedings of coling 2012* (pp. 2669–2684). Mumbai, India: The COLING 2012 Organizing Committee. Retrieved from `http://www.aclweb.org/anthology/C12-1163`

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.

Baayen, R. H. (2008). *Analyzing linguistic data* (1st ed.). Cambridge University Press. Retrieved from `http://gen.lib.rus.ec/book/index.php?md5=479AAB617AE91EFB8A3D7E6A6378890D`

Behaghel, O. (1932). *Deutsche syntax: eine geschichtliche darstellung. band iv. wortstellung. periodenbau.* Germany: Heidelberg: Carl Universitatsbuchhandlung.

Bell, A. (1984, 6). Language style as audience design. *Language in Society*, *13*, 145–204. Retrieved from `http://journals.cambridge.org/article_S004740450001037X` doi: 10.1017/S004740450001037X

Bock, J. K. (1982). Towards a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review*, 1–47.

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

Bock, J. K., & Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, *21*, 47–67.

Bock, K., Irwin, D., & Davidson, D. J. (2004). Putting first things first. In J. Henderson & F. Ferreira (Eds.), *The interface of language, vision, and action: Eye movements and the visual world* (pp. 249–278). Psychology Press.

Bornkessel, I., Schlesewsky, M., & Friederici, A. D. (2002). Grammar overrides frequency: evidence from the online processing of flexible word order. *Cognition*, *85*(2), B21 - B30. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0010027702000768` doi: http://dx.doi.org/10.1016/S0010-0277(02)00076-8

Boston, M. F., Hale, J. T., Patil, U., Kliegl, R., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*(1), 1–12. Retrieved from http://www.ling.uni-potsdam.de/~vasishth/Papers/jemrsurprisal.pdf

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*(2), B13 - B25. Retrieved from http://www.sciencedirect.com/science/article/pii/S0010027799000815 doi: http://dx.doi.org/10.1016/S0010-0277(99)00081-5

Breslow, N., & Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*(421), 9-25. Retrieved from http://www.jstor.org/stable/2290687

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the Dative Alternation. *Cognitive Foundations of Interpretation*, 69–94.

Chang, F., Dell, G. S., & Bock, K. (2006, April). Becoming Syntactic. *Psychological Review*, *113*(2), 234–272. Retrieved from http://dx.doi.org/10.1037/0033-295x.113.2.234

Chater, N., & Christiansen, M. H. (2010). Language evolution as cultural evolution: how language is shaped by the brain. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 623–628. Retrieved from http://dx.doi.org/10.1002/wcs.85 doi: 10.1002/wcs.85

Choi, H.-w. (2007). Length and Order: A Corpus Study of Korean Dative-Accusative Construction. *Discourse and Cognition*, *14*(3), 207–227.

Chomsky, N., & Miller, G. A. (1963). Introduction to the formal analysis of natural languages. In R. D. Luce, R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 2, pp. 269–322). New York: Wiley.

Clark, H. H., & Haviland, S. E. (1977). Comprehension and the Given-New Contract. In R. O. Freedle (Ed.), *Discourse Production and Comprehension* (pp. 1–40). Hillsdale, N. J.: Ablex Publishing.

Collins, M. (2012). *Cognitive perspectives on english word order* (Unpublished doctoral dissertation). The Ohio State University. (unpublished thesis)

Culicover, P. (2014). Constructions, complexity, and word order variation. In F. Newmeyer & L. Preston (Eds.), *Measuring grammatical complexity* (pp. 148–178). United Kingdom: Oxford University Press. Retrieved from http://www.zora.uzh.ch/84672/

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as

evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210. Retrieved from http://scholar.google.com/scholar.bib?q=info:1ulLoWI1IDoJ:scholar.google.com/&output=citation&hl=de&as_sdt=0,5&ct=citation&cd=0

Demberg, V., Keller, F., & Koller, A. (2013). Incremental, predictive parsing with psycholinguistically motivated tree-adjoining grammar. *Computational Linguistics*, 1025–1066.

Ferreira, V. S. (1996). Is it better to give than to donate? syntactic flexibility in language production. *Journal of Memory and Language*, *35*, 724–755.

Ferreira, V. S. (2003). The persistence of optional complementizer production: Why saying that is not saying that at all. *Journal of Memory and Language*, *48*(2), 379 - 398. Retrieved from http://www.sciencedirect.com/science/article/pii/S0749596X02005235 doi: http://dx.doi.org/10.1016/S0749-596X(02)00523-5

Ferreira, V. S., & Dell, G. S. (2000, June). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive psychology*, *40*(4), 296–340. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10888342

Ferrer i Cancho, R. (2004, Nov). Euclidean distance between syntactically linked words. *Phys. Rev. E*, *70*, 056135. Retrieved from http://link.aps.org/doi/10.1103/PhysRevE.70.056135 doi: 10.1103/PhysRevE.70.056135

Filippova, K., & Strube, M. (2007, June). Generating constituent order in German clauses. In *Acl 2007, proceedings of the 45th annual meeting of the association for computational linguistics*. Prague, Czech Republic: The Association for Computer Linguistics.

Francis, W. N., & Kučera, H. (1989). *Manual of information to accompany a standard corpus of present-day edited american english, for use with digital computers*. Brown University, Department of Linguistics.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336-10341. Retrieved from http://www.pnas.org/content/112/33/10336.abstract doi: 10.1073/pnas.1502134112

Gallo, C. G., Jaeger, T. F., & Smyth, R. (2008). Incremental syntactic planning across clauses. In *In proceedings of the 30th annual meeting of the cognitive science society* (pp. 845–850).

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependen-

cies. *Cognition*, *68*, 1–76.

Gibson, E. (2000). Dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain: Papers from the first mind articulation project symposium.* Cambridge, MA: MIT Press. Retrieved from http://www.ling.uni-potsdam.de/~vasishth/Papers/Gibson-Cognition2000.pdf

Gildea, D., & Temperley, D. (2007). Optimizing grammars for minimum dependency length. In *Proceedings of the 45th annual conference of the association for computational linguistics (acl-07)* (pp. 184–191). Prague. Retrieved from http://www.cs.rochester.edu/~gildea/pubs/gildea-temperley-acl07.pdf

Gildea, D., & Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science*, *34*(2), 286–310.

Gries, S. T. (2005). Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research*, *34*(4), 365–399. Retrieved from http://dx.doi.org/10.1007/s10936-005-6139-3 doi: 10.1007/s10936-005-6139-3

Gries, S. T., & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on alternations. *International Journal of Corpus Linguistics*, *9*(1), 97–129.

Gulordava, K., & Merlo, P. (2015, August). Diachronic trends in word order freedom and dependency length in dependency-annotated corpora of latin and ancient greek. In *Proceedings of the third international conference on dependency linguistics (depling 2015)* (pp. 121–130). Uppsala, Sweden: Uppsala University, Uppsala, Sweden. Retrieved from http://www.aclweb.org/anthology/W15-2115

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the north american chapter of the association for computational linguistics on language technologies* (pp. 1–8). Pittsburgh, Pennsylvania: Association for Computational Linguistics. Retrieved from http://dx.doi.org/10.3115/1073336.1073357 doi: 10.3115/1073336.1073357

Hawkins, J. A. (1994). *A Performance theory of order and constituency.* New York: Cambridge University Press.

Hawkins, J. A. (2000). The relative order of prepositional phrases in english: Going beyond manner-place-time. *Language Variation and Change*, *11*(03), 231–266. Retrieved from http://dx.doi.org/10.1017/S0954394599113012 doi: 10.1017/S0954394599113012

Hawkins, J. A. (2001). Why are categories adjacent? *Journal of Linguistics*,

37, 1–34.

Hawkins, J. A. (2003). Why are zero-marked phrases close to their heads? In G. Rohdenburg & B. Mondorf (Eds.), *Determinants of grammatical variation in english.* Berlin: De Gruyter Mouton.

Hawkins, J. A. (2004). *Efficiency and complexity in grammars.* Oxford University Press.

Hawkins, J. A. (2011). Discontinuous dependencies in corpus selections: Particle verbs and their relevance for current issues in language processing. In E. M. Bender & J. E. Arnold (Eds.), *Language from a cognitive perspective: Grammar, usage, and processing* (p. 269-290). CSLI Publishers.

Hawkins, J. A. (2014). *Cross-linguistic variation and efficiency.* Oxford University Press.

Jaeger, T. F. (2006). *Redundancy and syntactic reduction in spontaneous speech* (Unpublished doctoral dissertation). Stanford University.

Jaeger, T. F. (2010, August). Redundancy and reduction: Speakers manage information density. *Cognitive Psychology*, *61*(1), 23–62. Retrieved from `http://dx.doi.org/10.1016/j.cogpsych.2010.02.002`

Jaeger, T. F. (2011). Corpus-based research on language production: Information density and reducible subject relatives. In E. M. Bender & J. E. Arnold (Eds.), *Language from a cognitive perspective: Grammar, usage, and processing* (pp. 161–197). CSLI Publishers.

Jaeger, T. F., & Buz, E. (in press). Signal reduction and linguistic encoding. In E. M. Fernandez & H. S. Cairns (Eds.), *Handbook of psycholinguistics* (p. To appear). Wiley-Blackwell.

Jaeger, T. F., & Norcliffe, E. (2009). The cross-linguistic study of sentence production: State of the art and a call for action. *Language and Linguistic Compass*, *3*(4), 866–887. Retrieved from `http://dx.doi.org/10.1111/j.1749-818X.2009.00147.x`

Jaeger, T. F., & Tily, H. (2011). Language processing complexity and communicative efficiency. *WIRE: Cognitive Science*, *2*(3), 323–335. doi: 10.1002/wcs.126

James, F. (2000). *Modified kneser-ney smoothing of n-gram models* (Tech. Rep.). Moffett Field, CA, United States: RIACS.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth acm sigkdd international conference on knowledge discovery and data mining* (pp. 133–142). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/775047.775067` doi: 10.1145/775047.775067

Johansson, R., & Nugues, P. (2007, May). Extended constituent-to-

dependency conversion for English. In *Proceedings of nodalida 2007*. Tartu, Estonia. Retrieved from http://dspace.utlib.ee/dspace/bitstream/10062/2560/1/reg-Johansson-10.pdf

Konieczny, L. (2000, November). Locality and parsing complexity. *Journal of Psycholinguists Research*, *29*(6), 627–645. Retrieved from http://view.ncbi.nlm.nih.gov/pubmed/11196066

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32-59. doi: 10.1080/23273798.2015.1102299

Lee, M.-W., & Gibbons, J. (2007). Rhythmic alternation and the optional complementiser in English: new evidence of phonological influence on grammatical encoding. *Cognition*, *105*(2), 446–56. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17097626 doi: 10.1016/j.cognition.2006.09.013

Levelt, W., & Maasen, B. (1981). Crossing the boundaries in linguistics: Studies presented to manfred bierwisch. In W. Klein & W. Levelt (Eds.), (pp. 221–252). Dordrecht: Springer Netherlands. Retrieved from http://dx.doi.org/10.1007/978-94-009-8453-0_12 doi: 10.1007/978-94-009-8453-0_12

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126 - 1177. Retrieved from http://www.sciencedirect.com/science/article/pii/S0010027707001436 doi: http://dx.doi.org/10.1016/j.cognition.2007.05.006

Levy, R., Fedorenko, E., & Gibson, E. (2013). The syntactic complexity of russian relative clauses. *Journal of Memory and Language*, *69*(4), 461 - 495. Retrieved from http://www.sciencedirect.com/science/article/pii/S0749596X12001209 doi: http://dx.doi.org/10.1016/j.jml.2012.10.005

Levy, R., & Gibson, E. (2013). Surprisal, the pdc, and the primary locus of processing difficulty in relative clauses. *Frontiers in Psychology*, *4*(229).

Lewis, R. L., Vasishth, S., & Van Dyke, J. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, *10*(10), 447–454.

Linzen, T., & Jaeger, T. F. (2014). Investigating the role of entropy in sentence processing. In *Proceedings of the cognitive modeling and computational linguistics workshop at acl* (pp. 10–18). Baltimore, MD.

Linzen, T., & Jaeger, T. F. (2015). Uncertainty and expectation in sentence

processing: Evidence from subcategorization distributions. *Cognitive Science*, *40*(1).

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, *9*(2), 159-191. Retrieved from http://www.lingviko.net/JCS.pdf

Lohse, B., Hawkins, J. A., & Wasow, T. (2004). Domain Minimization in English Verb-Particle Constructions. *Language*, *80*(2), 238–261.

MacDonald, M. C. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes*, *9*(2), 157-201. Retrieved from http://lcnl.wisc.edu/publications/archive/132.pdf

MacDonald, M. C. (1999). Distributional information in language comprehension, production, and acquisition: Three puzzles and a moral. In B. MacWhinney (Ed.), *The emergence of language* (p. 177-196). Erlbaum. Retrieved from http://lcnl.wisc.edu/publications/archive/85.pdf

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*(226), 1-16. Retrieved from http://lcnl.wisc.edu/publications/archive/266.pdf (Published with commentaries in Frontiers.)

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994, 10). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*(4), 676-703. Retrieved from http://lcnl.wisc.edu/publications/archive/7.pdf

Maienborn, C. (2001). On the position and interpretation of locative modifiers. *Natural Language Semantics*, *9*(2), 191—240.

Manning, C. D. (2003). Probabilistic syntax. In R. Bod, J. B. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics.* Cambridge: MIT Press. Retrieved from get-book.cfm?BookID=5979

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993, June). Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, *19*(2), 313–330. Retrieved from http://dl.acm.org/citation.cfm?id=972470.972475

Martin, R. C., & Slevc, L. R. (2014). Language production and working memory. In M. Goldrick, V. S. Ferreira, & M. Miozzo (Eds.), *The oxford handbook of language production.* Oxford University Press.

Miller, T., & Schuler, W. (2008). A syntactic time-series model for parsing fluent and disfluent speech. In *Proceedings of the 22nd international conference on computational linguistics - volume 1* (pp. 569–576). Stroudsburg, PA, USA: Association for Computational Linguis-

tics. Retrieved from http://dl.acm.org/citation.cfm?id=1599081.1599153

Neumann, G., & van Noord, G. (1992). Self-monitoring with reversible grammars. In *Proceedings of the 14th conference on computational linguistics - volume 2* (pp. 700–706). Nantes, France: Association for Computational Linguistics. Retrieved from http://dx.doi.org/10.3115/992133.992178 doi: 10.3115/992133.992178

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics*, *31*(1).

Parker, R., Graff, D., Kong, J., Chen, K., & Maeda, K. (2011). English gigaword fifth edition. In *Linguistic data consortium.*

Petrov, S., Barrett, L., Thibaux, R., & Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics* (pp. 433–440). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dx.doi.org/10.3115/1220175.1220230 doi: 10.3115/1220175.1220230

Phillips, C. (2013). Some arguments and non-arguments for reductionist accounts of syntactic phenomena. *Language and Cognitive Processes*, *28*, 156-187. Retrieved from http://ling.umd.edu/~colin/wordpress/wp-content/uploads/2014/08/phillips2013_reductionism.pdf

Pickering, M. J., & Branigan, H. P. (1998, November). The Representation of Verbs: Evidence from Syntactic Priming in Language Production. *Journal of Memory and Language*, *39*(4), 633–651. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0749596X9892592X doi: 10.1006/jmla.1998.2592

Pickering, M. J., & Garrod, S. (2013, 8). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*, 329–347. Retrieved from http://journals.cambridge.org/article_S0140525X12001495 doi: 10.1017/S0140525X12001495

Pickering, M. J., & Traxler, M. J. (2003). Evidence against the use of subcategorisation frequency in the processing of unbounded dependencies. *Language and Cognitive Processes*, *18*(4), 469-503.

Pollard, C., & Sag, I. (1994). *Head-Driven Phrase Structure Grammar.* University Of Chicago Press.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., & Webber, B. (2008). The penn discourse treebank 2.0. In *Proceedings of the sixth international conference on language resources and evalu-*

ation (lrec'08). Marrakech, Morocco: European Language Resources Association (ELRA).

Qian, T., & Jaeger, T. F. (2012). Cue effectiveness in communicatively efficient discourse production. *Cognitive Science*, *36*(7), 1312–1336. Retrieved from http://dx.doi.org/10.1111/j.1551-6709.2012.01256.x doi: 10.1111/j.1551-6709.2012.01256.x

Rajkumar, R., & White, M. (2014). Better surface realization through psycholinguistics. *Language and Linguistics Compass*, *8*(10), 428–448. Retrieved from http://dx.doi.org/10.1111/lnc3.12090 (ISSN: 1749-818X) doi: 10.1111/lnc3.12090

Reitter, D., Keller, F., & Moore, J. D. (2011). A computational cognitive model of syntactic priming. *Cognitive Science*, *35*(4), 587–637. Retrieved from http://www.david-reitter.com/pub/reitter2011syntacticpriming.pdf doi: 10.1111/j.1551-6709.2010.01165.x

Reitter, D., & Moore, J. D. (2014). Alignment and task success in spoken dialogue. *Journal of Memory and Language*, *76*, 29–46. Retrieved from http://www.david-reitter.com/pub/reitter2014JML-alignment.pdf doi: 10.1016/j.jml.2014.05.008

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 1 - volume 1* (pp. 324–333). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1699510.1699553

Roland, D., Elman, J. L., & Ferreira, V. S. (2006). Why is that? structural prediction and ambiguity resolution in a very large corpus of english sentences. *Cognition*, *98*(3), 245 - 272. Retrieved from http://www.sciencedirect.com/science/article/pii/S0010027705000028 doi: http://dx.doi.org/10.1016/j.cognition.2004.11.008

Ros, I., Santesteban, M., Fukumura, K., & Laka, I. (2015). Aiming at shorter dependencies: the role of agreement morphology. *Language, Cognition and Neuroscience*, *30*(9), 1156-1174. doi: 10.1080/23273798.2014.994009

Schuler, W. (2014). Sentence processing in a vectorial model of working memory. In *Fifth annual workshop on cognitive modeling and computational linguistics (CMCL 2014)*.

Schuler, W., AbdelRahman, S., Miller, T., & Schwartz, L. (2010, March).

Broad-coverage parsing using human-like memory constraints. *Computational Linguistics*, *36*, 1–30. Retrieved from http://dx.doi.org/10.1162/coli.2010.36.1.36100 doi: http://dx.doi.org/10.1162/coli.2010.36.1.36100

Scontras, G., Badecker, W., Shank, L., Lim, E., & Fedorenko, E. (2015). Syntactic complexity effects in sentence production. *Cognitive Science*, *39*(3), 559–583. Retrieved from http://dx.doi.org/10.1111/cogs.12168 doi: 10.1111/cogs.12168

Shain, C., van Schijndel, M., Gibson, E., & Schuler, W. (2016, March). Exploring memory and processing through a gold standard annotation of Dundee. In *Proceedings of cuny 2016.* Gainesville, Florida, USA: University of Florida.

Slevc, L. R. (2011). Saying what's on your mind: working memory effects on sentence production. *Journal of experimental psychology. Learning, memory, and cognition*, *37*(6), 1503–1514. Retrieved from http://dx.doi.org/10.1037/a0024350

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302319.

Snider, N. (2009). Similarity and structural priming. In *Proceedings of the 31th annual conference of the cognitive science society* (pp. 815–820).

Snider, N., & Zaenen, A. (2006). Animacy and syntactic structure: Fronted nps in english. In M. Butt, M. Dalrymple, & T. King (Eds.), *Intelligent linguistic architectures: Variations on themes by ronald m. kaplan.* Stanford: CSLI Publications.

Stallings, L. M., MacDonald, M. C., & O'Seaghdha, P. G. (1998, 10). Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-np shift. *Journal of Memory and Language*, *39*(3), 392–417. Retrieved from http://lcnl.wisc.edu/publications/archive/16.pdf

Staub, A., Clifton, & Frazier, L. (2006). Heavy NP shift is the parser's last resort: Evidence from eye movements. *Journal of Memory and Language*, *54*(3), 389–406+. Retrieved from http://www.sciencedirect.com/science/article/B6WK4-4J5T5VK-1/2/354774d1fe4312802b4723c88b4aefab

Szmrecsanyi, B. (2004). On Operationalizing Syntactic Complexity. In G. a. Purnelle, C. a. Fairon, & A. Dister (Eds.), *Le poids des mots. proceedings of the 7th international conference on textual data statistical analysis. louvain-la-neuve, march 10-12, 2004* (Vol. II, pp. 1032–1039). Louvain-la-Neuve: Presses universitaires de Louvain.

Temperley, D. (2003). Ambiguity avoidance in english relative clauses.

Language, *79*(3), 464–484.

Temperley, D. (2007). Minimization of dependency length in written English. *Cognition*, *105*(2), 300–333. Retrieved from http://www.sciencedirect.com/science/article/B6T24-4M7CDMS-2/2/e095449f6439b30003822a5838e53786 doi: DOI:10.1016/j.cognition.2006.09.011

Tily, H. (2010). *The role of processing complexity in word order variation and change* (Unpublished doctoral dissertation). Stanford University. (unpublished thesis)

Traxler, M. J., Pickering, M. J., & Clifton, C. (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language*, *39*(4), 558-592+.

Trueswell, J., Tanenhaus, M., & Garnsey, S. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, *33*(3), 285 - 318. Retrieved from http://www.sciencedirect.com/science/article/pii/S0749596X8471014X doi: http://dx.doi.org/10.1006/jmla.1994.1014

Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints in sentence processing: separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(3), 528.

van Schijndel, M., Exley, A., & Schuler, W. (2013). A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, *5*(3), 522–540.

van Schijndel, M., Nguyen, L., & Schuler, W. (2013, August). An analysis of memory-based processing costs using incremental deep syntactic dependency parsing. In *Proceedings of cmcl 2013.* Sofia, Bulgaria: Association for Computational Linguistics.

van Schijndel, M., & Schuler, W. (2013, June). An analysis of frequency- and memory-based processing costs. In *Proceedings of naacl-hlt 2012.* Atlanta, Georgia, USA: Association for Computational Linguistics.

van Schijndel, M., Schuler, W., & Culicover, P. W. (2014, July). Frequency effects in the processing of unbounded dependencies. In *Proceedings of cogsci 2014.* Quebec, Quebec, Canada: Cognitive Science Society.

Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, *82*(4), 767–794. Retrieved from http://www.ling.uni-potsdam.de/~vasishth/Papers/Vasishth-Lewis-Language2006.pdf

Warren, T., & Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition*, *85*, 79-112.

Wasow, T. (2002). *Postverbal behavior*. Stanford: CSLI Publications. Retrieved from `get-book.cfm?BookID=3678`

Wasow, T., & Arnold, J. (2003). *Post-verbal constituent ordering in english*. Mouton.

Webber, B. (2004). D-LTAG: Extending lexicalized TAG to discourse. *Cognitive Science*, *28*(5), 751-779.

Webber, B. (2006). Accounting for discourse relations: Constituency and dependency. In *Intelligent linguistic architectures* (pp. 339–360). CSLI Publications.

Webber, B., Stone, M., Joshi, A., & Knott, A. (2003). Anaphora and discourse structure. *Computational Linguistics*, *29*(4). Retrieved from `http://www.aclweb.org/anthology/J03-4002.pdf`

White, M., & Rajkumar, R. (2012, July). Minimal dependency length in realization ranking. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 244–255). Jeju Island, Korea: Association for Computational Linguistics. Retrieved from `http://www.aclweb.org/anthology/D12-1023`

Wiechmann, D., & Lohmann, A. (2013, 3). Domain minimization and beyond: Modeling prepositional phrase ordering. *Language Variation and Change*, *25*, 65–88. Retrieved from `http://journals.cambridge.org/article_S0954394512000233` doi: 10.1017/S0954394512000233

Wu, S., Bachrach, A., Cardenas, C., & Schuler, W. (2010). Complexity metrics in an incremental right-corner parser. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 1189–1198). Uppsala, Sweden: Association for Computational Linguistics. Retrieved from `http://dl.acm.org/citation.cfm?id=1858681.1858802`

Yamashita, H., & Chang, F. (2001). "Long before short" preference in the production of a head-final language. *Cognition*, *81*.

Yngve, V. H. (1960). A Model and an Hypothesis for Language Structure. *Proceedings of the American Philosophical Society*, *104*(5), 444–466.

Predictor	Entity counted	Brown Acc%	WSJ Acc%
Gibson's definition	Discourse referents	69.95	67.95
Temperley's definition	Non-punctuation words	69.67; $p = 0.49$	66.91; $p = 1.03e-05$
Syllable-based definition	Stressed syllables	69.29; $p = 0.25$	66.77; $p = 0.46$

Table A.1: Individual classification accuracies of various definitions of dependency length on Brown (8385 data points) and WSJ corpora (20330 data points), with statistical significance determined using McNemar's χ-square test against the previous row

Appendix A. Supplementary Analyses and Figures

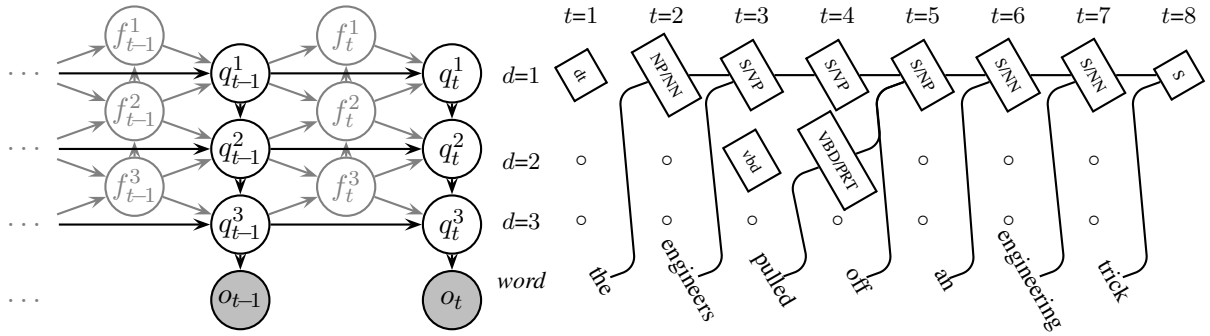
8.1 Ranking Experiments: Which dependency length measure predicts syntactic choice best?

In the literature, dependency length calculations have been defined in multiple ways. Our first experiment compares three common definitions of dependency length—those of Gibson (2000), Temperley (2007) and Anttila et al. (2010)—in order to ascertain the most effective definition of dependency length for predicting syntactic choice. Gibson's DLT formulation measures dependency length in terms of the number of intervening discourse referents (nouns and verbs). Temperley (2007) measures dependency length by counting the number of words between heads and dependents (punctuation marks are excluded and adjacent words are accorded a distance of 1). Anttila et al. (2010) provide a prosodic definition of dependency length whereby head-dependent distances are counted in terms of the number of intervening stressed syllables (see Section 3 for further details). To calculate dependency length, each dataset consisting of constituent structure syntactic trees (corresponding to both reference and variant sentences) is first converted to a corpus of dependency trees using the LTH constituency to dependency converter[21] (Johansson & Nugues 2007) and head-dependent distances corresponding to each definition above are calculated.

We evaluate the accuracy of each dependency length measure described above in choosing the corpus sentence over the generated variants in our datasets (Brown and WSJ corpora). Ties are resolved by choosing one alternative randomly and then averaging results across 10 runs. Gibson's discourse referent-based definition of dependency length outperforms the other two definitions in terms of absolute ranking accuracy with both corpora (see Table A.1). Note, however, that the ranking accuracy of Gibson's definition is significantly higher than Temperley's word-based definition only

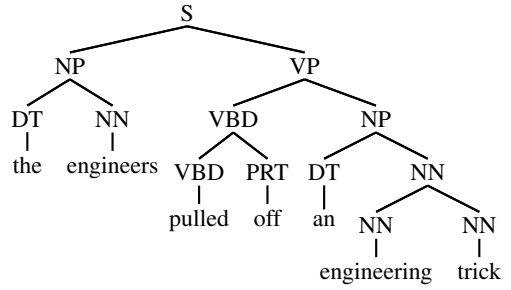
[21] `http://nlp.cs.lth.se/software/treebank-converter`

for the WSJ corpus. Syllable-based dependency length (Anttila's definition) performs worse than the other two definitions for both corpora. For both datasets, dependency length measured in terms of words gives same trends of results for regression and classifation results reported in Tables 6 and 5 respectively using dependency length in number of discourse referents reported as conclusions of this paper.

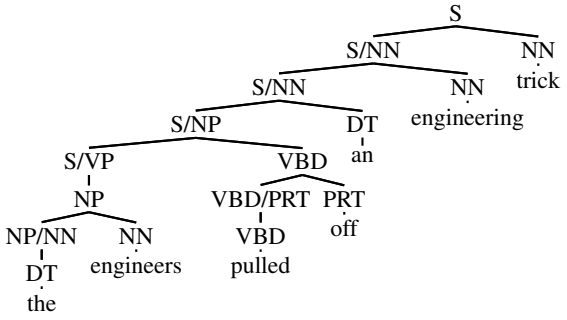


(a) Dependency structure in the HHMM parser. Conditional probabilities at a node are dependent on incoming arcs.

(b) HHMM parser as a store whose elements at each time step are listed vertically, showing a good hypothesis on a sample sentence out of many kept in parallel. Variables corresponding to q_t^d are shown.



(c) A sample sentence in CNF.



(d) The right-corner transformed version of (c).

Figure A.1: Reproduced from Wu et al. (2010): Various graphical representations of HHMM parser operation. (a) shows probabilistic dependencies. (b) considers the q_t^d store to be incremental syntactic information. (c)–(d) demonstrate the right-corner transform, similar to a left-to-right traversal of (c). In 'NP/NN' we say that NP is the *active* constituent and NN is the *awaited*.

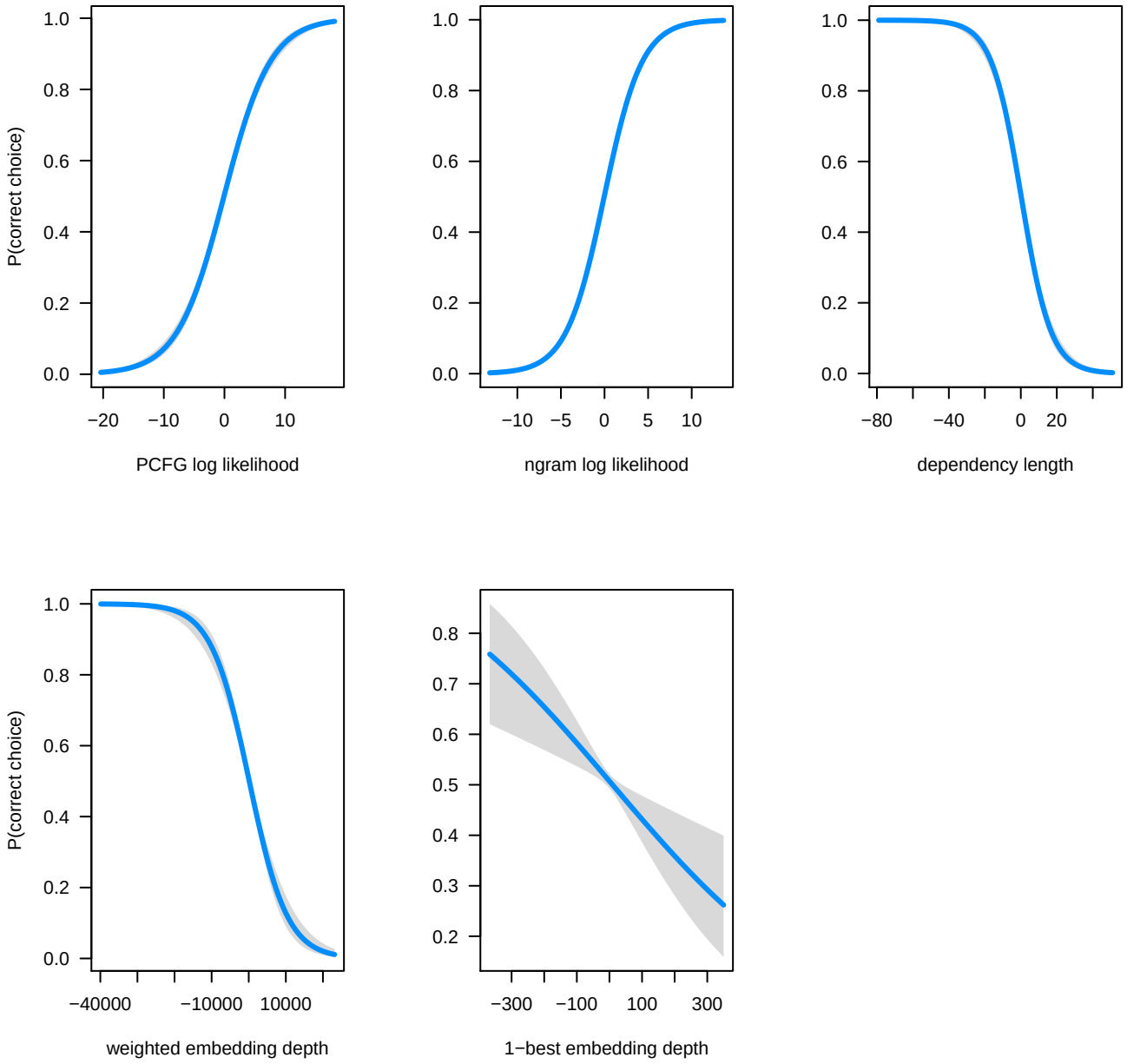


Figure A.2: Effects plot of all predictors in the full model for the Brown corpus (gray band shows confidence interval)

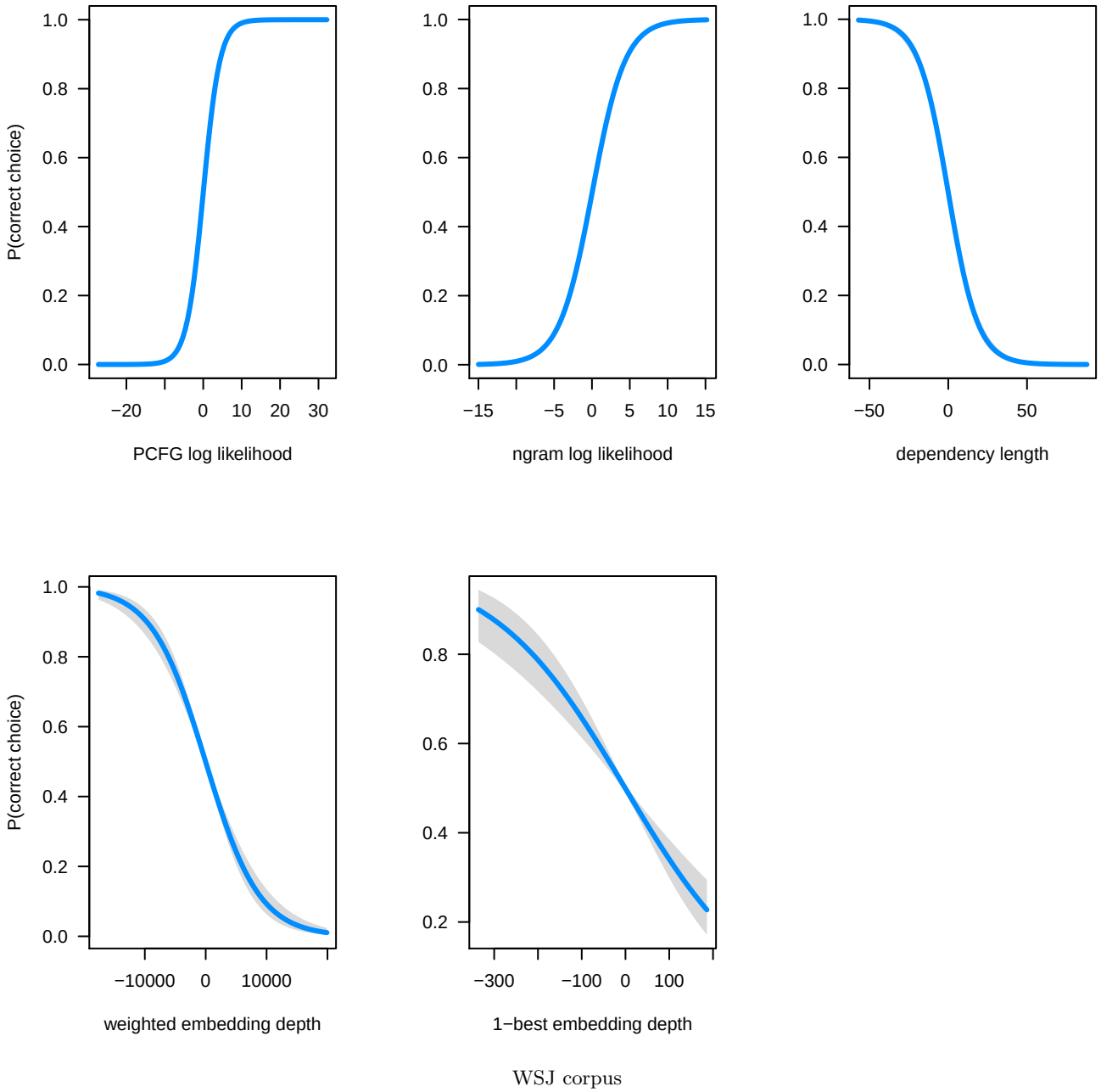
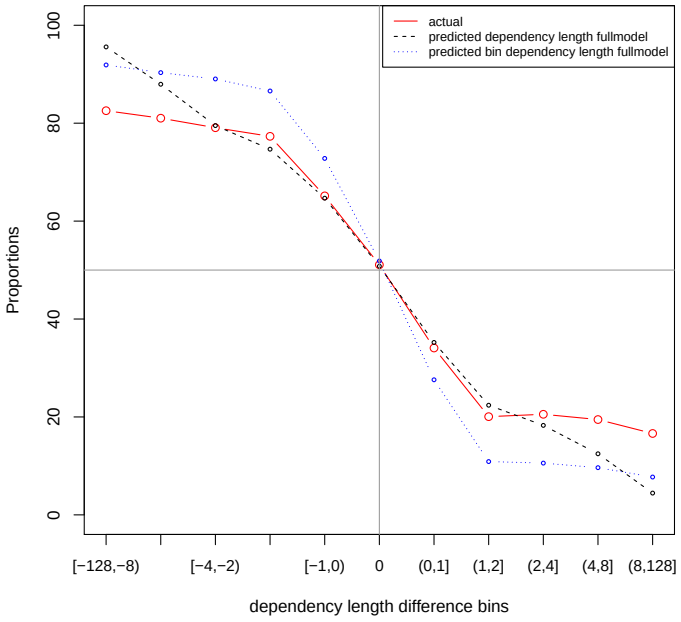
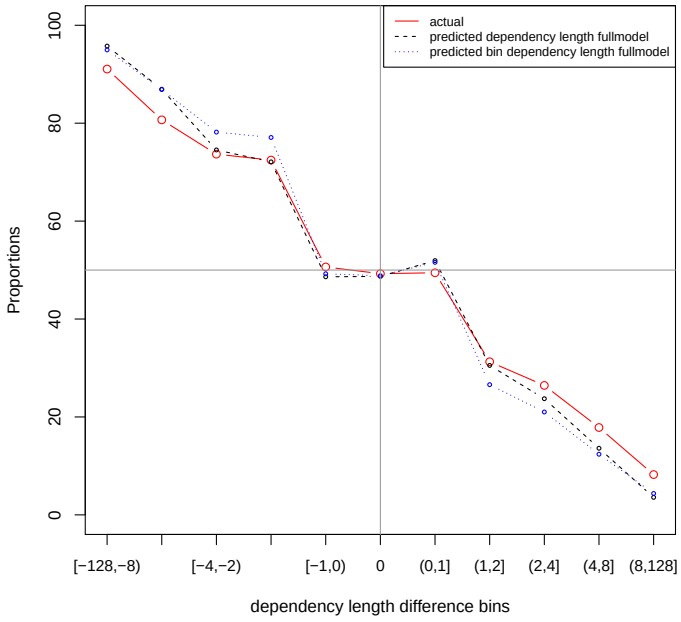


Figure A.3: Effects plot of all predictors in the full model for the WSJ corpus (gray band shows confidence interval)



a Brown corpus



b WSJ corpus

Figure A.4: Correct choice proportions of actual data and full models containing dependency length and binned dependency length respectively

Appendix B. Supplementary Data

Our data files (which serve as input to the statistical analyses scripts written in R) have been made publicly available via the open source data repository, Dataverse. The data can be downloaded by via the link: `http://dx.doi.org/10.7910/DVN/1RUSDZ`

Highlights

- We show that integration costs stipulated by Dependency Locality Theory are indeed a significant predictor of syntactic choice in written English even in the presence of competing frequency-based and cognitively motivated control factors including surprisal.
- The predictions of dependency length and surprisal are only moderately correlated, a finding which mirrors results for sentence comprehension.
- The efficacy of dependency length in predicting the corpus choice increases with increasing head-dependent distances.
- The tendency towards dependency minimization is reversed in some cases and surprisal is effective in these non-locality cases.