DEPENDENCY LOCALITY AS AN EXPLANATORY PRINCIPLE FOR WORD ORDER

RICHARD FUTRELL
University of California, Irvine

ROGER P. LEVY
Massachusetts Institute of Technology

EDWARD GIBSON
Massachusetts Institute of Technology

This work focuses on explaining both grammatical universals of word order and quantitative word-order preferences in usage by means of a simple efficiency principle: dependency locality. In its simplest form, dependency locality holds that words linked in a syntactic dependency (any head–dependent relationship) should be close in linear order. We give large-scale corpus evidence that dependency locality predicts word order in both grammar and usage, beyond what would be expected from independently motivated principles, and demonstrate a means for dissociating grammar and usage in corpus studies. Finally, we discuss previously undocumented variation in dependency length and how it correlates with other linguistic features such as head direction, providing a rich set of explananda for future linguistic theories.*

Keywords: word-order universals, efficiency, corpus studies, processing, dependency-length minimization

1. INTRODUCTION. A growing consensus in functional linguistics is that the universals of human language are best explained in terms of pressures having to do with communicative efficiency (von der Gabelentz 1901, Zipf 1936, 1949, Hockett 1960, Slobin 1973, Givón 1991, Hawkins 1994, 2004, 2014, Chomsky 2005, Christiansen & Chater 2008, Jaeger & Tily 2011, Fedzechkina et al. 2012, Gibson et al. 2019). The idea is that languages are shaped by a trade-off between information transfer and ease of production and understanding under the information-processing constraints inherent to the human brain. We refer to this idea as the EFFICIENCY HYPOTHESIS. The current article focuses on explaining both grammatical universals of word order and quantitative word-order preferences by means of a simple efficiency principle: dependency locality. We give large-scale evidence that dependency locality predicts word order in both grammar and usage, showing a means for dissociating the two in corpus studies. Also, we discuss previously undocumented variation in dependency length and how it correlates with head direction, providing a rich set of explananda for future linguistic theories.

While the efficiency hypothesis makes predictions about grammar, it requires the study of quantitative properties of usage, because efficiency in communication is fundamentally about the things that speakers actually say. In this article we distinguish grammar and usage in the following way. By GRAMMAR we mean a conventional formal system linking phonological form, on the one hand, and some representation of meaning, on the other. This mapping can be expressed equivalently as a function from form to meaning or from meaning to form, or in terms of an intermediate representation from which both meaning and form are derived.[1] By USAGE we mean the probability distribution over utterances used by a community at some time, of which a corpus is a large sample.[2]

* We thank Jack Hawkins, Kristina Gulordava, and audiences at EMNLP 2015, the University of Edinburgh Center for Language Evolution Colloquium, and the UC Irvine Quantitative Approaches to Language Science conference for helpful feedback. We thank two anonymous referees whose comments improved the article. This work was supported in part by NSF Linguistics DDRI Grant #1551543 to the first author. All errors are our own.

[1] For example, the standard EST/Y-model in generative syntax (Chomsky & Lasnik 1977, Chomsky 2000, 2005:10–11, Irurtzun 2009).

[2] It is possible to talk coherently about probabilistic properties of usage without implying any stochasticity or meaningfully quantitative properties of grammar. The stochastic nature of usage may come entirely from

The usage distribution determines efficiency because it states which utterances and utterance parts are used more frequently. Under the simplest notion of communicative efficiency, these parts should be easier to express. This idea follows from the general principle that more frequent messages should take less effort to transmit, a key result from information theory (Shannon 1948, Cover & Thomas 2006:110–12).

We study the phenomenon of dependency locality using large-scale analysis of parsed corpora of naturalistic text. While the large-scale approach allows us to make a highly general claim, it does not supplant careful analysis of individual examples, the usual methodology in formal linguistics. In the case of dependency locality, such studies have already been carried out for a variety of cases, as we review in §2.2; the large-scale analysis provides evidence that the effect is likely to hold in detailed analysis of languages and constructions not yet examined. We believe a large-scale corpus-analytic approach complements more traditional linguistic analysis, by providing macroscale characterizations of phenomena that may be instantiated in a variety of ways at the microscale, and by verifying that the patterns discovered in individual constructions and languages hold more generally across constructions and in utterances containing many interacting constructions.

However, this is not only a corpus study. In the section on grammar and usage we compare observed orders of utterances to possible orders for those utterances as estimated using a probabilistic model of grammatical word orders. Our work therefore introduces a new methodology to corpus linguistics: studying not only distributions in corpora, but also the properties of controlled models of grammar induced from those corpora. This method allows us to make arguments about both grammar and usage based on usage data.

This article is primarily a study of word order. Modern generative approaches to syntax, where 'narrow syntax' is defined as an intermediate representation between form and meaning, have often disavowed word order as a true property of syntax, considering it part of the syntax-sensorimotor interface. In this view, word order is a property of phonological form only, resulting from the linearization of an unordered hierarchical syntactic representation in the course of 'externalization' (Kayne 1994, Chomsky 2007). If we adopt this view, our results are consistent with the idea that communicative optimization happens in externalization, while abstract syntactic competence may be determined by arbitrary and possibly innate computational constraints (but for arguments that core properties of narrow syntax, in particular recursion, can arise from communicative need, see Piantadosi & Fedorenko 2017). We do not dispute this view here directly. But the ultimate aim of our research program is to eventually explain as much as possible about human language in terms of the efficiency hypothesis.

2. Background: dependency locality. Dependency locality is the idea that in grammar and usage there exists a pressure for words linked in a syntactic dependency to be close to each other. It is typically (but not always) stated in terms of dependency grammar, but this is not necessary for the core predictions. Below, we review the conceptual foundations of dependency locality, some of the typological patterns that it has been used to explain, and previous corpus evidence for it.

variation in what meanings speakers wish to convey. Thus the prepositional phrase *to New York* is much more probable in current English usage than *to Nichewaug*, not because this is a fact about grammar, but because people are more likely to want to express the meaning of the former than the latter. Stochasticity in grammar would mean that the mapping between form and meaning is probabilistic—which it may well be, but this conclusion is not established merely by observing that usage is probabilistic.

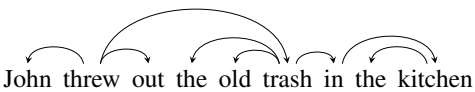
2.1. THE NOTION OF SYNTACTIC DEPENDENCY. This study is about dependency length: the distances between linguistic heads and dependents. The notions of head and dependent can be defined on top of most syntactic formalisms. Nearly all theories of syntax include some notion of headedness: the idea that the behavior of a constituent can be understood primarily by reference to one distinguished word, the HEAD (Bloomfield 1933, Tesnière 1959, Hays 1964, Bresnan 1982, Hudson 1984, Pollard & Sag 1987, Mel'čuk 1988, Corbett et al. 1993). For example, the syntactic behavior of a verb phrase is determined primarily by the head verb in the phrase. The property of headedness in syntax is also known as ENDOCENTRICITY. A DEPENDENT is a word that modifies a head, and a DEPENDENCY is the relationship between a head and a dependent.

While most constituents appear to be endocentric, not all syntactic formalisms posit a head for all phrases. In these formalisms, some constructions forming constituents are EXOCENTRIC, having no head. For example, it is notoriously difficult to assign a head to a coordinated phrase such as *Bob and Mary* (Temperley 2005, Popel et al. 2013), with different dependency formalisms choosing different means (Tesnière 1959, Mel'čuk 1988), and some introducing elements of phrase structure formalisms especially for this purpose (Hudson 1990). In syntactic formalisms such as are used in the MINIMALIST PROGRAM, all phrases have heads, but these heads may be silent elements (Adger 2003). Despite these difficulties, in the majority of cases, phrases are uncontroversially endocentric and heads can be identified (Corbett et al. 1993).

Taking the notion of endocentricity to its logical conclusion, DEPENDENCY GRAMMAR posits that syntax can be fully described solely in terms of relationships among head words, without a further notion of constituent or other higher-order groupings of words (Tesnière 1959, Hays 1964, Hudson 1984, Mel'čuk 1988, Sleator & Temperley 1991). In dependency grammar, a sufficient syntactic analysis takes the form of a tree (or graph) linking heads to their dependents. If all phrases are endocentric, then constituency grammars and dependency grammars can be freely converted one to the other.

Examples of dependency trees are given in 1. The verb *throw* is the head of two nouns that modify it, *John*—its subject—and *trash*—its object. Subject and object relations are kinds of dependency relations.[3]

(1)



Another way to think about dependency is to note that heads and dependents are words that must be linked together in order to understand a sentence, to a first approximation. For example, in order to correctly understand sentence 1, a comprehender must determine that a relationship of adjectival modification exists between the words *old* and *trash*, and not between, say, the words *old* and *kitchen*. In typical syntactic dependency analyses, objects of prepositions (*him* in *for him*) depend on their prepositions, articles depend on the nouns they modify, and so on.

In this work, while we use a dependency formalism, we do not wish to claim that dependency grammar is the only correct description of syntax, nor that a dependency tree encapsulates all of the syntactic information that there is to know about a sentence. We only wish to claim that dependency trees represent an important and large subset of that information. We describe syntax in terms of dependency trees here for two reasons: simplicity and convenience. With regard to simplicity, dependency trees are simple to

[3] The parse represents a case where *in the kitchen* is taken to depend on *trash*, not on *threw*: thus the parse represents one possible resolution of a PP attachment ambiguity.

reason about and to formulate algorithms over, while still providing a strong description of syntactic structure. With regard to convenience, large-scale corpora are available with dependency annotation, because the natural language processing community has discovered that it is easier to perform this annotation in a consistent way across languages than to use a constituency annotation (Nivre 2015).

2.2. WHAT IS DEPENDENCY LOCALITY? The concept of dependency locality is that words linked in syntactic dependencies should be close to each other. We can operationalize this idea as the DEPENDENCY-LENGTH MINIMIZATION (DLM) hypothesis: that language users prefer word orders that minimize total DEPENDENCY LENGTH per sentence: where dependency length is the linear distance between words in dependency relations. The hypothesis makes two broad predictions. First, when the grammar of a language provides multiple ways to express an idea, language users will prefer the expression with the shortest total dependency length. Second, grammars should enforce word-order rules that enable short dependencies in usage. For recent reviews on the history of and evidence for this idea, see Dyer 2017, Liu et al. 2017, and Temperley & Gildea 2018.

The idea of DLM has been proposed in various forms for almost a century. We believe the oldest statement of the idea is by Behaghel (1930:30–31), who proposed two relevant laws of word order.

(i) OBERSTES GESETZ ('highest law'): That which belongs together mentally is placed close together.
(ii) GESETZ DER WACHSENDEN GLIEDER ('law of growing constituents'): Of two sentence components, the shorter goes before the longer, when possible.

The HIGHEST LAW can be operationalized as DLM: Behaghel's examples are that adjectives modifying nouns are close to those nouns, adverbs modifying adjectives are close to those adjectives, and so on. We now understand that the LAW OF GROWING CONSTITUENTS (originally formulated in Behaghel 1909) arises as a corollary of DLM in head-initial contexts, as we discuss below in §2.3.

A principle closely related to DLM was proposed as an underlying explanation for word-order universals by Rijkhoff (1986:98–99) (as the 'principle of head proximity') and Hawkins (1990), and the specific quantitative formulation in terms of distance between words linked in dependency grammar was formulated by Hudson (1995) and first applied to dependency corpora by Ferrer-i-Cancho (2004).

An issue that arises in the definition of DLM is how to measure distance between heads and dependents. A common approach has been to measure the distance in terms of the number of intervening words (Heringer et al. 1980, Hudson 1995, Wasow 2002). Other proposals have included the number of intervening discourse referents (Gibson 1998), the number of syllables (Benor & Levy 2006), the number of lexical stresses (Anttila et al. 2010), and the complexity of the intervening material (Chomsky 1975: 477, Wasow 2002). In practice, the proposed measures are highly correlated with each other (Wasow 2002, Shih & Grafmiller 2011).

The functional motivation for DLM comes from the idea that short dependencies make comprehension and production more efficient: that languages are structured to enable maximal information transfer with minimal effort, where long dependencies are argued to incur greater processing effort for various reasons. Theoretical proposals differ in terms of WHY they posit that long dependencies incur greater processing effort. Essentially these explanations can be divided into two categories: those based on time constraints and those based on memory constraints.

Hawkins (1994) makes a proposal based on time constraints. He proposes that a pressure very similar to DLM (which he calls EARLY IMMEDIATE CONSTITUENTS; EIC) arises because it minimizes the search time required for a parser to determine the correct head of a phrase. Under the assumption that the incremental parser, when encountering a word, searches for the head of that word linearly outward from the word, the search time is minimized when the head is close to the current word.

A more common explanation for the difficulty of long dependencies relies on notions of limited memory resources available in parsing or generation. The intuition is that dependency length corresponds to the amount of time a word representation must be kept in working memory during sentence processing, and this amount of time corresponds to difficulties in memory retrieval. To see this, consider a case where two words are linked in a dependency, such as *threw* and *out* from 1. We know that they are linked in a dependency because understanding the word *out* in context requires that it be combined with the word *threw* to form the phrasal verb *threw out*. In incremental parsing, when the parser reaches the word *out*, it must integrate a representation of the word *out* with a representation of the previous word *threw* based on its memory of the context leading up to *out*. If the dependency between *threw* and *out* is long, then the representation of the context word *threw* will have been in working memory for a long time, during which time the representation will have been subject to progressive decay and cumulative interference. The result is that retrieving the representation may be difficult or inaccurate, and more difficult and more inaccurate the longer the representation has been in memory. A similar story can be told from the language generation side: by the time a speaker is preparing to produce *out*, she may have forgotten which exact words she said previously, requiring a working-memory retrieval operation to know that the context contained the word *threw*.

In psycholinguistics, the idea that long dependencies correspond to human parsing difficulty due to working-memory pressures is represented most prominently in the form of the DEPENDENCY LOCALITY THEORY (Gibson 1998, 2000). Similar proposals were made by Just and Carpenter (1992) and Hudson (1995). In this theory, when words in dependencies are separated by a large number of new discourse referents, processing slowdown results, and this slowdown is called a DEPENDENCY LOCALITY EFFECT. Gibson (1998) shows that dependency locality effects can be seen as a common phenomenon underlying the difficulty of understanding multiply center-embedded clauses and object-extracted relative clauses. Dependency locality effects subsume effects of stack depth in parsing, which were studied by Yngve (1960). Grodner and Gibson (2005) show reading-time evidence for dependency locality effects in sentences such as 2a–c.

(2) a. The **administrator** who the nurse **supervised** …
b. The **administrator** who the nurse from the clinic **supervised** …
c. The **administrator** who the nurse who was from the clinic **supervised** …

In these sentences, the dependency between *administrator* and *supervised* is progressively lengthened, and a resultant reading-time increase at and after the word *supervised* is observed (for further experimental evidence, see Bartek et al. 2011).

While dependency locality effects are well attested and robust in controlled experiments, they are not the whole story when it comes to language processing difficulty. In addition to dependency locality effects there also exist ANTILOCALITY EFFECTS (Konieczny 2000), where the processing time at a final verb in a verb-final language decreases when more material is placed before the verb. Considerations such as these have led to the development of SURPRISAL THEORY (Hale 2001, Levy 2008), a qualita-

tively different theory of processing difficulty in which the difficulty associated with a word is not a function of working-memory retrievals, but rather a function of the probability of the word in context. On its own, surprisal theory does not predict locality effects (Levy 2008:1139–41), but locality effects are predicted by a recent extension of surprisal theory based on the idea of prediction from noisy memory representations (Futrell & Levy 2017, Futrell 2019).

Another, more general motivation for dependency locality comes from the theory of statistical complexity (Crutchfield & Young 1989, Shalizi & Crutchfield 2001). To our knowledge, this motivation for dependency locality has not been discussed in the linguistic literature before. Statistical complexity theory characterizes the complexity of any sequence of symbols—such as a linguistic utterance—in terms of the minimal information required to predict the future of the sequence accurately given the past of the sequence up to some point. Complex sequences require more information about the past, and simple sequences require less. The quantity of information about the past that is useful for predicting the future is called EXCESS ENTROPY (Shalizi & Crutchfield 2001:848–50), and its application to natural language has been studied in detail by Dębowski (2011). It turns out that excess entropy increases whenever elements in a sequence that are statistically dependent on each other are separated from each other by a large distance, such that excess entropy is lower when dependent words are close to each other—closely related to the principle of dependency locality (Futrell 2019). Therefore, under the assumption that human language is constrained to be simple in the statistical-complexity sense, it should exhibit dependency locality.

For now, we focus on showing large-scale evidence for DLM as an empirical principle driving quantitative patterns of word order. We do not take a position on the precise functional motivation for DLM: we believe such an argument would be best served by detailed experimental studies, rather than the relatively coarse-grained approach we take here. We assume that the dependencies whose length is minimized are syntactic dependencies as defined by dependency grammar, and that the correct distance metric for dependency length is the number of intervening words; these decisions are driven by expedience, and the fact that the predictions of the theory do not change substantially across formalisms and distance metrics.

We also emphasize that dependency locality is a PRESSURE affecting word order in grammar and usage; we do not claim that the word order in every utterance minimizes dependency length, or that dependency locality is the only principle that determines word order. We claim only that grammars and usage preferences are structured such that words in dependencies are typically close. Within any given language there may be individual constructions that violate dependency locality, but we claim that these constructions will be rare both within and across languages.

2.3. WHAT CAN DEPENDENCY LOCALITY EXPLAIN? In the previous literature, dependency locality has been used to explain (at least) four major properties of natural language word order. Here we survey these predictions and how they are derived from DLM, as well as other explanations for these phenomena.

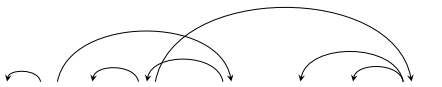
PROJECTIVITY AND THE CONTIGUITY OF CONSTITUENTS. Dependency locality provides a potential explanation for one of the most pervasive and theoretically important formal features of natural language word order: that it is typically PROJECTIVE, meaning that when dependency connections are drawn above a sentence, the lines do not cross. Stated in terms of dependency grammar, this property may seem arcane. However, it corresponds to one of the deepest ideas in syntax: that the order of semantic composition is isomorphic to word order (at least before movement). If we take the dependency

structure of a sentence to represent its order of semantic composition, then projectivity means that word order is ISOMORPHIC to the order of composition. For example, in a phrase such as *he read the book quickly*, the phrase *the book* is built up before it is combined as a unit with the verb phrase, and as such no word from elsewhere in the verb phrase can intervene between *the* and *book*, which would violate both isomorphism and projectivity by creating a crossing dependency. Sentences that violate this principle of isomorphism, such as **He read the quickly book*, are rarely grammatical across languages (but see the nonconfigurational languages for exceptions: e.g. Hale 1983). The idea that the order of semantic composition should be isomorphic to surface word order is considered a major structural principle of language (Culbertson & Adger 2014:5843).

Projectivity means that constituents correspond to contiguous sequences of words; nonprojectivity arises in cases of discontinuity (Groß & Osborne 2009). As such, projectivity in dependency grammar corresponds formally to CONTEXT-FREENESS in phrase structure grammar: the fact that constituents are typically nested inside each other and do not interleave (Marcus 1965:181). While natural languages are not strictly context-free (Shieber 1985), they deviate from context-freeness only rarely and with strong formal restrictions (Nivre & Nilsson 2005, Havelka 2007, Ferrer-i-Cancho et al. 2018, Yadav et al. 2019); the consensus among formal language theorists studying human languages is that natural languages are 'mildly context-sensitive' (Weir 1988, Joshi et al. 1991; cf. Michaelis & Kracht 1997, Bhatt & Joshi 2004, Kobele 2006), meaning that they go only slightly beyond context-free languages in terms of the Chomsky hierarchy (Chomsky 1959, Chomsky & Schützenberger 1963).

Nonprojective or non-context-free structures arise in language in the form of DISPLACEMENT phenomena. These comprise right extraposition (often a result of HEAVY NP SHIFT), as shown in sentence 3, where *who you know* modifies *woman*, and WH-movement, as shown in sentence 4, where *what* is the object of the verb *did*. In minimalist frameworks, projective structures are built by MERGE (or 'external merge') operations, and MOVE (or 'internal merge') operations serve to create potentially nonprojective structures (Chomsky 1995, 2004, Stabler 1997, Michaelis 1998, 2001). Constraints on movement, such as the PHASE IMPENETRABILITY CONDITION (Chomsky 2000:108), are closely related to constraints on nonprojectivity in dependency trees (Pitler et al. 2013:20–21).

(3)



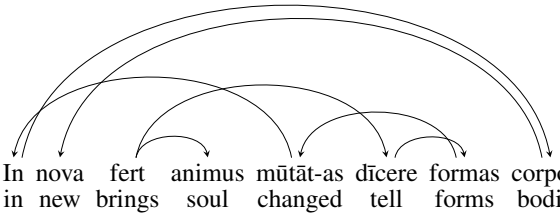
(4)

DLM has been advanced as an explanation for the apparent constraints on nonprojectivity in natural language. The formal observation that DLM could potentially explain the strong tendency to projectivity was first made by Ferrer-i-Cancho (2006), who noted empirically that minimizing dependency length results in trees with very few nonprojective arcs; in subsequent work, Ferrer-i-Cancho (2016) has shown analytically that minimization of dependency length leads to a reduction in the number of nonprojective arcs.[4]

[4] Rijkhoff (1990:9) similarly claimed that the 'principle of domain integrity' (i.e. projectivity) follows from Behaghel's first law, but considered the 'principle of head proximity' (i.e. DLM) to be a separate principle.

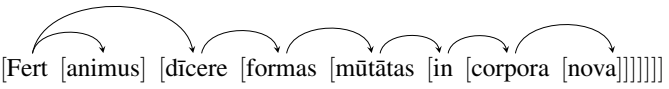
As an example, consider sentence 5a from Ovid, which has a highly nonprojective word order in its original form. A random projective linearization of the same sentence is shown as sentence 5b; it has observably lower dependency length (and consists of nested constituents, consistent with the identification of projectivity with context-freeness).

(5) a.



'My soul brings me to tell of forms changed into new bodies.'

b.



The sentence appears in highly nonprojective form in the original, with very high dependency length, because the poet is operating under metrical constraints. We do not know which word order would have been easiest to produce and comprehend in Latin, but the DLM hypothesis leads to the conjecture that word order in colloquial Latin would have had substantially lower dependency length, and fewer deviations from projectivity as a result.

Dependency locality thus provides a simple and general explanation for one of the most characteristic features of natural language syntax: the principle that constituents are contiguous, or equivalently that the order of semantic composition is isomorphic to surface word order, or that languages are overwhelmingly context-free. Yet DLM is more general and makes somewhat different predictions from principles such as these.

While DLM favors projective word orders, it is not the case that minimal-dependency-length word orders are necessarily projective. It is in some cases possible to achieve lower dependency lengths by breaking projectivity (Chung 1984, Hochberg & Stallmann 2003, Park & Levy 2009). Indeed, deviations from projectivity such as right extraposition in English appear to serve the purpose of dependency locality, inasmuch as they are a form of heavy NP shift (Newmeyer 2014:300). It remains to be seen whether the deviations from projectivity in natural language serve to strategically lower dependency length more generally; this is a promising avenue of research (Ferrer-i-Cancho & Gómez-Rodríguez 2016).

Instead of positing that the underlying factor explaining word order is a constraint for a certain kind of isomorphism between form and meaning, dependency locality posits that the underlying factor is that words that must be integrated together syntactically will be close to each other. In some cases, this factor works against projectivity. In fact, we show in §4 that dependency length in corpora is even shorter than we would expect from projectivity as an independent principle.

While there is a compelling case to be made for dependency locality as the explanatory factor behind the overwhelming projectivity of natural language, there are other possible explanations for this formal property. One possible alternative functional explanation is from computational (time) complexity: projective dependency grammar enables faster parsing than nonprojective dependency grammar. Technically, exhaustively parsing a sentence with a projective dependency grammar takes time on the order $O(n^3)$, where n is the number of words in a sentence, whereas exhaustively parsing with a nonprojective dependency grammar generally has worse time complexity and can even be NP-hard (Kuhlmann 2013:371–73).

SHORT-BEFORE-LONG AND LONG-BEFORE-SHORT CONSTITUENT ORDERING PREFERENCES. There is a strong tendency in predominantly head-initial languages such as English to order constituents after a head from short to long (Behaghel 1909, Wasow 2002). There is overwhelming corpus evidence for this tendency in constructions where word-order variation is possible in English (Wasow 2002, Bresnan et al. 2007, Shih et al. 2015). The pattern is also reflected in heavy NP shift, a phenomenon where a long NP constituent is moved to appear later in a sentence than would otherwise be grammatical, as shown in English in sentences 6a–d.

(6) a. He ate [the popsicle] quickly.
b. *He ate quickly [the popsicle].
c. He ate [the popsicle that he had bought after spending all day in the sun] quickly.
d. He ate quickly [the popsicle that he had bought after spending all day in the sun].

In these examples, the grammar of English typically disallows an adverb intervening between a verb and its object, but allows an exception when the object is very long, creating a short-before-long order.[5] DLM predicts this pattern in head-initial contexts, as noted by Hawkins (1994) and Temperley (2007). The logic is demonstrated in 7, which shows how short-before-long order is advantageous in this setting.

(7) a.


b.


While DLM predicts short-before-long orders in head-initial contexts, it predicts a long-before-short preference in head-final contexts, as shown in 8. Correspondingly, a long-before-short preference has been demonstrated experimentally in Japanese, an overwhelmingly head-final language (Yamashita & Chang 2001), and heavy NP shift in Japanese appears to move heavy elements to the left rather than to the right (Hawkins 1994, Chang 2009). An emergent long-before-short preference is also observable in head-final languages in artificial language learning studies (Fedzechkina et al. 2017).

(8) a.


b.


CONSISTENCY IN HEAD DIRECTION. In contexts where heads have small numbers of dependents, DLM predicts that dependencies in a language should be either consistently

[5] An interesting question is why the grammar of English prefers objects to be closer to the verb in examples such as 6a–b, even though 6b has shorter dependency length than 6a. In this connection, we note that the dependency length of 6b is only very slightly longer than 6a, perhaps not enough to make an appreciable difference in terms of processing efficiency. Furthermore, extensions of the theory of DLM, such as information locality (Futrell 2019), predict that certain dependencies will be under stronger pressure to be short than others; if the verb–object dependency is under stronger minimization pressure than the verb–adverb dependency, then the order verb–object–adverb could be preferred even at the cost of a slight increase in gross dependency length.

head-final or consistently head-initial. The reasoning for this prediction is demonstrated in 9, for tree structures where each head has only one dependent. Consistency in head direction within languages forms the basis of some of the most well-known universals of word order, the HARMONIC WORD-ORDER CORRELATIONS (Greenberg 1963, Vennemann 1974, Dryer 1992), comprising Greenberg's universals 2 through 6.

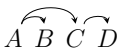
(9) a. 

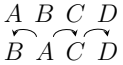
b. 

Consistency in head direction has been the subject of explanations other than DLM. It motivated the idea of a 'head direction parameter' in the PRINCIPLES-AND-PARAMETERS approach to syntax (Chomsky 1981, Baker 2001:68). More generally, consistency in head direction can be motivated in terms of simplicity in grammars. If a grammar has consistent head direction, then a learner does not need to learn a separate parameter describing the branching direction of every head; a single parameter suffices, and the grammar is simpler and easier to learn (Hsu & Chater 2010). However, we show in §4 that dependency length in corpora is even shorter than one would expect from consistency in head direction alone, so it may turn out not to be necessary to posit consistency in head direction as an independent factor in order to explain crosslinguistic patterns in word order.

EXCEPTIONS TO CONSISTENCY IN HEAD DIRECTION. While head direction is typically consistent within languages, there are often exceptions. In particular, these exceptions are typically for very short or one-word constituents: for example, in Spanish, which is highly head-initial, determiners come before nouns (where determiners are typically considered dependents of nouns in dependency grammar and DLM studies). These exceptions are documented by Dryer (1992).

Precisely these exceptions are predicted by DLM in contexts where heads have many dependents. Example 10 shows how inconsistency in head direction for short constituents can result in lower dependency length than a strategy of purely consistent head direction. In this example, *A* has two dependents, and minimal dependency length is achieved by placing these two dependents on opposite sides of the head. This prediction of DLM was derived by Gildea and Temperley (2007, 2010) and presaged by work on similar problems in abstract graph theory (Hochberg & Stallmann 2003). Gildea and Temperley (2007) found that the optimal strategy to achieve minimal dependency length (while maintaining projectivity) is to place dependents on alternating sides of their head outward in order of increasing length.

(10) a. 

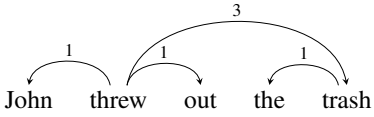
b. 

2.4. PREVIOUS CORPUS EVIDENCE FOR DEPENDENCY-LENGTH MINIMIZATION. Quantitative corpus evidence has always formed a major part of the support for dependency locality, starting with Behaghel's (1909) study of quantitative word-order patterns in German, Greek, and Latin. For reviews of recent evidence, see Liu et al. 2017 and Temperley & Gildea 2018. Here we focus on the previous studies that are most relevant to our current work.

A common approach has been to compare dependency length in attested sentences to dependency length under various random baselines. Dependency length per sentence is typically measured as the sum of the distance from each head to each dependent, with distance measured as the number of intervening words, as shown in 11. This approach

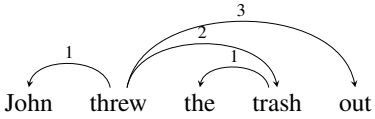
tells us not only whether words in dependencies are close, but also whether they are closer than we would expect given alternative hypotheses. The choice of random baseline defines the precise hypothesis about dependency length being tested.

(11) a.



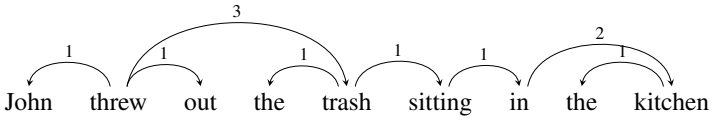
Total dependency length = 6

b.



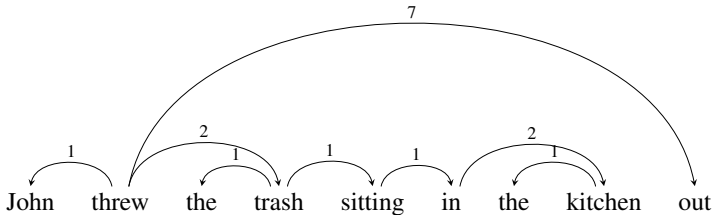
Total dependency length = 7

c.



Total dependency length = 11

d.

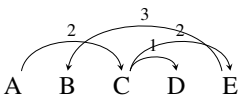


Total dependency length = 16

Studies using random baselines fall broadly into three categories. We call these random trees, random orders, and grammatical orders.

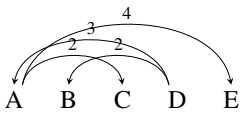
RANDOM-TREE studies such as Liu 2008 compare dependency length in a real sentence to dependency length in random trees with the same number of nodes. For a corpus sentence five words in length, this baseline would compare dependency length in the real sentence to dependency length in baseline sentences, such as those shown in 12.

(12) a.



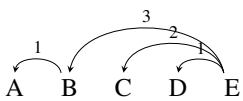
Total dependency length = 8

b.



Total dependency length = 11

c.

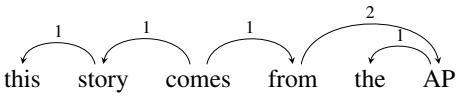


Total dependency length = 7

RANDOM-ORDER baselines compare dependency length in real sentences to dependency lengths in random reorderings of the words of those sentences, while holding the tree topology constant, as exemplified in 13. Works taking this approach include Park & Levy 2009, Gildea & Temperley 2010, Futrell, Mahowald, & Gibson 2015, Gildea &

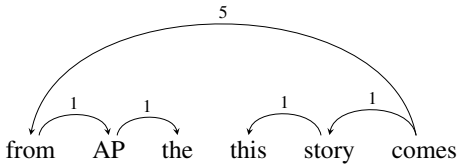
Jaeger 2015, Dyer 2017, Gómez-Rodríguez et al. 2019, and Yu et al. 2019. Random-order baselines tell us whether dependency length is shorter than we would expect if the constraints operative in word order were those embodied by the random baseline. For example, if a random baseline works by reordering words in a sentence while maintaining a constraint of head-finality, and if natural language has dependency length less than this baseline, then this tells us that natural language dependency length is minimized beyond what we would expect if head-finality were the only operative constraint in syntactic linearization.

(13) a.



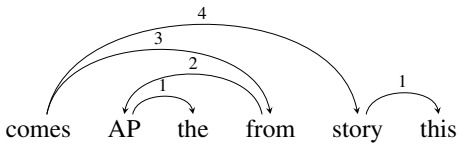
Total observed dependency length = 6

b.



Total baseline dependency length = 9

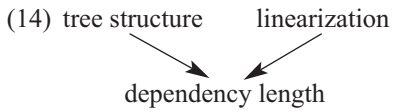
c.



Total baseline dependency length = 11

Random-tree and random-order baselines test subtly but importantly different variants of the DLM hypothesis. The ultimate dependency length of a sentence is a function of two variables: the dependency tree structure of the sentence—regardless of information about the linear order of words—and the rules and preferences by which that tree is linearized. This situation is schematized in 14. Random-tree baselines vary both the tree structure and the linearization rules, and show that some combination of the choice of tree structures and word-order rules in natural language result in lower dependency length than the baseline. In contrast, random-tree baselines hold tree structure constant and vary linearization rules, and show us that specifically word-order rules and preferences are shaped by a pressure for DLM.

(14)



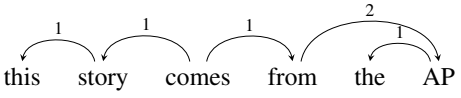
The interpretation of random-tree baselines is complex. To some extent, the dependency tree structure of a sentence, considered without regard to the linear order of the words, can be seen as a representation of the meaning being expressed by the utterance. This is especially true when the dependency arcs are labeled with grammatical functions such as SUBJECT and OBJECT. Thus when we find that attested sentences have lower dependency length than the random-tree baseline, this minimization might be due not to word-order rules and preferences, but rather to speakers selecting which meanings to express such that the resulting utterances have short dependency length. For example, words and clauses may be dropped in order to lower dependency length as compared to a random-tree baseline. It would also be favorable to use sentences with fewer adjuncts, because trees with more dependents per node will tend to have longer

dependency length (Esteban et al. 2016). A random-tree baseline cannot distinguish between DLM accomplished by word-order rules and DLM accomplished by a usage preference for propositions which, when linearized according to the rules of the language, result in low dependency length.

Random trees and random orders have also been used as baselines to study other linguistic phenomena. For example, Ferrer-i-Cancho et al. 2018 studies rates of crossing dependencies in a random-order baseline, Courtin & Yan 2019 studies frequencies of various subtree configurations in real trees as opposed to random trees and random orders, and Yadav et al. 2019 compares the gap-degree properties of real and random trees.

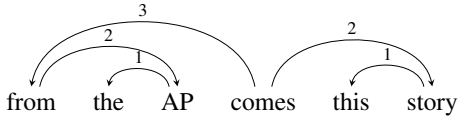
The third kind of baseline is what we call GRAMMATICAL ORDERS. These baselines compare dependency length in real sentences to dependency length in alternative orders for those sentences under the grammar of the language. An example is shown in 15. This is the approach most commonly taken in more detailed corpus studies of specific constructions, such as Hawkins 1998, Wasow 2002, and Rajkumar et al. 2016, where it is possible for the experimenters to generate by hand the grammatically possible alternative orders for sentences. Grammatical-orders baselines test the hypothesis that language users select orders with minimal dependency length when grammar provides them with multiple options; it shows an effect of DLM in usage preferences, not necessarily in grammatical rules. We present a large-scale application of the grammatical-orders baseline, using probabilistic models of word order to approximate the set of alternative grammatical orders possible for an utterance, and show how these grammatical-orders baselines can be used to argue for DLM in grammar as well as in usage preferences.

(15) a.



Total observed dependency length = 6

b.



Total baseline dependency length = 9

In this work, we first extend the results of Futrell, Mahowald, & Gibson 2015, which used only random-order baselines, to more baselines and more languages (§4). Next, we present new corpus studies dissociating grammar and usage with respect to DLM using grammatical-orders baselines (§5). Next, we present an in-depth analysis of crosslinguistic variation in dependency length (§6).

3. DATA SOURCES. The data source for our corpus studies is the Universal Dependencies project, release 2.1 (Nivre et al. 2017). Here we discuss the nature of this data and the filters and transformations that we applied to it. We are making our complete data-processing pipeline available online.[6]

3.1. BACKGROUND ON UNIVERSAL DEPENDENCIES. Universal Dependencies (UD) is a collaborative project to create dependency-parsed corpora of many languages following a unified standard formalism. Its primary purpose is to create resources for training and testing parsers. Data annotation is accomplished either by hand or by hand-correction of automatic parses. The original parses are sometimes done according to the UD stan-

[6] http://github.com/langprocgroup/cliqs/

dard, and sometimes according to another standard, which is subsequently transformed automatically to the UD standard.

The genres of the underlying texts vary from language to language, but most languages have data predominantly from newspapers, blogs, fiction and nonfiction literature, and Wikipedia. Some corpora also include spoken language. Some corpora are of classical, literary, and liturgical languages, including Latin, Ancient Greek, Old Church Slavonic, and Gothic. In the case of Latin and Ancient Greek, the corpora include a great deal of metrical poetry. The corpora of Old Church Slavonic and Gothic consist of religious texts. The precise data sources for each corpus can be viewed online at the UD website.[7]

Table 1 lists the languages and corpora available in UD 2.1 that we use in this study. We exclude corpora that contain fewer than 500 sentences.[8] We also exclude the corpus of Telugu, which is based on examples from a grammar text, and not from naturalistic language production. This is not a perfectly typologically balanced sample: Indo-European languages predominate. However, there is a good amount of diversity in the languages, with samples of languages from twelve families, including Dravidian, Turkic, and Afroasiatic languages. Projects are under way to increase the linguistic diversity of UD, including in-progress corpora of Georgian, Somali, and Yoruba.

LANGUAGE	FAMILY (SUBFAMILY)	LANGUAGE	FAMILY (SUBFAMILY)
Latvian	Indo-European (Baltic)	Church Slavonic	Indo-European (Slavic)
Irish	Indo-European (Celtic)	Croatian	Indo-European (Slavic)
Afrikaans	Indo-European (Germanic)	Polish	Indo-European (Slavic)
Danish	Indo-European (Germanic)	Russian	Indo-European (Slavic)
English	Indo-European (Germanic)	Serbian	Indo-European (Slavic)
German	Indo-European (Germanic)	Slovak	Indo-European (Slavic)
Gothic	Indo-European (Germanic)	Slovenian	Indo-European (Slavic)
Dutch	Indo-European (Germanic)	Ukrainian	Indo-European (Slavic)
Norwegian (Bokmål)	Indo-European (Germanic)	Upper Sorbian	Indo-European (Slavic)
Norwegian (Nynorsk)	Indo-European (Germanic)	Arabic	Afroasiatic (Semitic)
Swedish	Indo-European (Germanic)	Hebrew	Afroasiatic (Semitic)
Ancient Greek	Indo-European (Greek)	Vietnamese	Austroasiatic (Viet)
Modern Greek	Indo-European (Greek)	Indonesian	Austronesian (Austronesian)
Bengali	Indo-European (Indo-Aryan)	Basque	Isolate (Basque)
Hindi	Indo-European (Indo-Aryan)	Tamil	Dravidian (Dravidian)
Northern Kurdish	Indo-European (Iranian)	Telugu	Dravidian (Dravidian)
Persian	Indo-European (Iranian)	Japanese	Japonic (Japanese)
Latin	Indo-European (Italic)	Korean	Koreanic (Korean)
Catalan	Indo-European (Romance)	Buryat	Mongolic (Mongolic)
French	Indo-European (Romance)	Mandarin	Sino-Tibetan (Sinitic)
Galician	Indo-European (Romance)	Uyghur	Turkic (Karluk)
Italian	Indo-European (Romance)	Kazakh	Turkic (Kipchak)
Portuguese	Indo-European (Romance)	Turkish	Turkic (Oghuz)
Romanian	Indo-European (Romance)	Estonian	Uralic (Finnic)
Spanish	Indo-European (Romance)	Finnish	Uralic (Finnic)
Bulgarian	Indo-European (Slavic)	Northern Sami	Uralic (Sami)
Czech	Indo-European (Slavic)	Hungarian	Uralic (Ugric)

TABLE 1. Languages with more than 500 sentences available in Universal Dependencies v2.1.

[7] http://universaldependencies.org

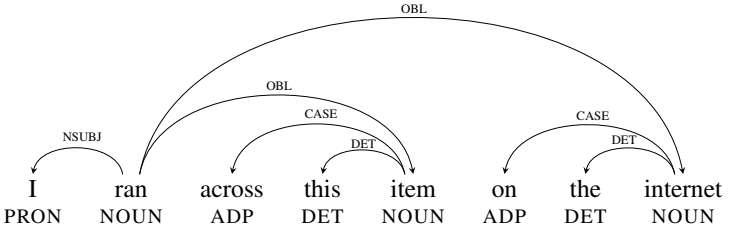
[8] This results in the exclusion of corpora of Belarusian, Cantonese, Coptic, Lithuanian, Marathi, and Sanskrit.

UD corpora contain part-of-speech tags and lemmas for each word, as well as a dependency tree over each sentence. Every word has exactly one head, and each dependency is labeled with a DEPENDENCY TYPE, such as NSUBJ (nominative subject), OBJ (direct object), AMOD (adjectival modification), and so on.

3.2. LINGUISTIC QUALITY OF UNIVERSAL DEPENDENCIES. The UD project is ambitious in proposing a unified syntactic standard for all languages. In pursuit of this ideal, it has been necessary to make compromises and trade-offs to make sure that similar structures across languages are annotated in the same way. For a review of the UD standard from the perspective of linguistic typology, see Croft et al. 2017, which finds that the dependency structures of UD are generally linguistically adequate, including some of their more controversial decisions. For an alternative perspective, see Osborne & Gerdes 2019. Here we survey some aspects of the UD standard that are relevant for dependency-length studies.

The most controversial aspect of UD annotation is what is called 'content-head' annotation style. This means that the primary dependencies annotated are between content words, with function words as dependents of those content words, rather than the other way around. For example, 16 shows an example parse from the UD English corpus where the prepositional phrase *on the internet* is parsed such that *internet* is the head of *on*. The UD standard works this way in order to allow locative phrases to be parsed in a uniform way across languages. In languages such as Finnish where *on the internet* would be expressed using a single word—*internet* in the locative case—the head of the phrase would be the inflected word *internet*. Thus in English, the head of the phrase is also chosen to be *internet*, with *on* as a dependent with dependency type CASE.

(16)



Content-head annotation style has the following implications: nouns are considered heads of adpositions, predicates are considered heads of copulas, content verbs are considered heads of auxiliary verbs (e.g. in *I can see you*, *see* is the head of *can*), and verbs are considered heads of complementizers. Also, nouns are held to be heads of their modifying adjectives and determiners, against the DP hypothesis (Abney 1987, Alexiadou et al. 2007, Bruening 2009). For recent work developing an alternative standard to UD that reflects a more standard syntactic analysis, see Gerdes et al. 2018, 2019.

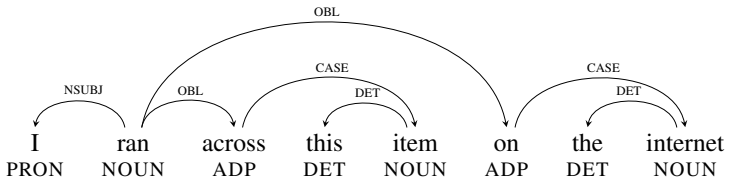
While content-head annotation style does not reflect the usually agreed-upon syntactic analysis, it leaves many major syntactic relations that are key to DLM unchanged. For example, the relations between a verb and its objects and adjuncts are analyzed uncontroversially.

3.3. FILTERS AND TRANSFORMATIONS. We do not use the UD corpora in raw form; rather we use them after applying several filters and transformations to make sure the data is interpretable in the way we want. One simple transformation is that we remove all punctuation from the corpora. Other transformations are more complex.

For studies of dependency length, we are interested in true syntactic dependencies. In order to obtain these dependencies, we applied several transformations to the UD cor-

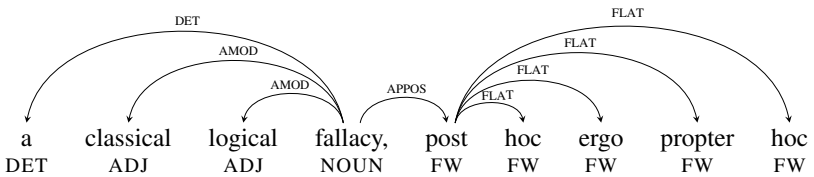
pora. For the studies reported in §4, corpora were transformed automatically so that adpositions are heads of their nouns in adpositional phrases, and copulas are heads of their predicates, and complementizers are heads of verbs.[9] In cases where a transformation was not possible, the offending sentence was removed from the corpus. Thus a sentence such as 16 is transformed to the structure shown in 17.

(17)



Another issue with UD annotation that could potentially affect dependency-length measurements is the treatment of word sequences that do not admit a natural dependency analysis. One example is foreign phrases. The basic parse structure for a foreign phrase in UD is shown in 18.

(18)



UD uses the 'flat' structure of 18 for foreign phrases, multiword names, compounds, fixed expressions such as *as well as*, lists, and coordinated phrases. In this flat structure, the first word is taken to be the head and all other words are taken to be dependents of it. This structure is used even in languages such as Japanese that are predominantly head-final in their true syntactic dependencies.

For foreign phrases, multiword names, and fixed expressions, we collapse the structure into one long word with no internal structure. This is because these structures do not admit any intervening material and are best thought of as unanalyzable units from the perspective of word order. For the remaining flat structures, we leave them as parsed by the UD standard.

In §4 we present results from corpora transformed in the manner described above. However, this transformation is not required to obtain our basic results on DLM. In the appendix, we present the same analyses with the original UD content-head dependencies, finding the same results. We also present analyses from corpora where all function words have been removed, thus removing potential confounds involving the treatment of function words.

3.4. Discussion. We have argued that UD provides an accurate and linguistically adequate representation of crosslinguistic syntax for the purposes of studying dependency length. Any annotation project of the magnitude of UD must make some compromises in the syntactic formalism it uses, pending the development of a syntactic formalism capable of uncontroversially describing all structures in all languages. We have argued that UD's compromises are sensible and conservative from the perspective of measuring dependency length.

[9] It is not possible in general to reverse the UD annotation to make auxiliary verbs into heads of content verbs, because content verbs are often assigned multiple dependent auxiliary verbs and it is not possible in general to recover the structure among these auxiliary verbs. Therefore we do not attempt this transformation.

One potential objection to the validity of any results we obtain is that the dependency parses as produced by the humans of the UD project may themselves contain a bias for short dependencies. In that case our results would show DLM in UD, but not necessarily in natural language. Based on the UD parsing guidelines, we do not believe they contain explicit biases in favor of short dependencies; nevertheless, we have attempted to address this kind of issue in two ways. First, we have replicated the results above using different dependency formalisms (see the appendix), suggesting that at least a subset of the UD parsing decisions are not necessary to show the DLM effect. Second, we have eliminated constructions from the corpus where the parsing standard could potentially bias the results, such as foreign phrases, as described above.

4. INDEPENDENT BASELINES. Here we present evidence that dependency length in natural language usage is shorter than we would expect from independently motivated constraints. We study the constraints of projectivity, consistency in head direction, and fixedness of word order with respect to syntactic dependency type. As noted previously, dependency locality has itself been advanced as a possible explanation for projectivity and consistency in head direction. But if these constraints alone sufficed to explain the observed dependency length in natural language and there were no DLM beyond them, then evidence for dependency locality as the causal factor would be weakened, since these constraints have reasonable independent motivations. If, however, dependency length is even shorter than we would expect from these constraints, then the idea of dependency locality as the single causal force behind these word-order patterns would be strengthened.

4.1. DEFINITION OF BASELINES. We employ random-reordering baselines as described and justified in §2.4. Our baselines instantiate three constraints, which can often be intersected, yielding a total of six individual random baselines. Each baseline describes what dependency length would be like if the given constraints were the only constraints operative in natural language linearization, across languages and utterances.

Our random baselines generate random linearizations of sentences. We measure dependency length in 100 random samples under each baseline, in order to approximate the expected dependency length under that baseline.

PROJECTIVE BASELINE. Projective baselines are generated by taking a sentence and reordering its words while maintaining the constraint that dependency lines do not cross. These baselines were also used in Gildea & Temperley 2010, Futrell, Mahowald, & Gibson 2015, and Dyer 2017. Dependency length under various mildly nonprojective baselines (subject to certain formal constraints on crossings) have been studied in Park & Levy 2009 and Gómez-Rodríguez et al. 2019; these baselines usually give higher average dependency length than a fully projective baseline.

CONSISTENT-HEAD-DIRECTION BASELINE. In these baselines, all linearizations are strictly head-final. The results would be equivalent for strictly head-initial linearizations: all that matters in terms of dependency length is that head direction is consistent. A baseline with projectivity and consistent head direction was presented in Futrell, Mahowald, & Gibson 2015; here we present consistent-head-direction baselines both with and without projectivity, and with and without fixed word order.

FIXED WORD ORDER. These baselines simulate grammars in which the linearization of a tree is a deterministic function of the tree structure and the dependency relation types. For example, in such a language, adjectives may consistently come before the nouns they modify. We generate random grammars of this type by assigning to each UD de-

pendency type a random weight in the interval [−1, 1]. Dependency types with negative weight are linearized head-finally, and dependency types with positive weight are linearized head-initially. Furthermore, when a head has multiple dependents on one side, the magnitude of the weight determines their order. Thus, for example, a dependency whose type gives it weight −0.9 will appear before a dependency whose type gives it weight −0.2, with both appearing before the head. This random grammar is similar to the one developed in Gildea & Temperley 2010, except that we assign weights to dependency types rather than to phrase labels.

Our fixed-word-order baseline requires projectivity. It is possible to develop algorithms for generating grammars with random fixed nonprojective word-order rules, but we expect that these would have higher average dependency length than the random fixed projective grammars, and hence would not materially change our present conclusions.

For the fixed-word-order baseline, we select one random grammar and linearize all of the sentences in a language according to it. The 100 samples from this baseline represent 100 random grammars, each applied uniformly to all of the sentences in a language. This method differs from Futrell, Mahowald, & Gibson 2015, where a new grammar was generated for each random sample of each sentence.

4.2. Results. We measure the dependency length of a real or reordered sentence as the sum of the lengths of each dependency in the sentences, where the length of a dependency is the number of intervening words between the head and the dependent plus one. For individual sentences, this measure is equivalent up to a constant factor to the metric of mean dependency distance proposed by Liu (2008).

Figure 1 shows mean dependency length as a function of sentence length for all languages and baselines.[10] By looking at mean dependency length as a function of sentence length, we control for variance across corpora in sentence length. We believe that variation in sentence length across corpora is likely mostly a function of genre and stylistic conventions, rather than any linguistically important variable.

In Fig. 1, the black line is the observed dependency length; the colored lines are baselines. We see that observed dependency length is regularly lower than any baseline dependency length. More precisely, we can speak of the growth rate of dependency length as a function of sentence length (as in Esteban et al. 2016): we find that observed dependency length grows more slowly than any baseline dependency length.

Statistical analysis. In order to quantify the DLM effect and assess its statistical significance, we fit regression models to predict dependency length from sentence length. For each baseline, we fit two regression models: the first using one slope to model both real and baseline data, and the second using one slope for the real data and another slope for the baseline data. If the latter model is a significantly better fit to the data than the former, then this means that the two slopes are significantly different; that is, the real dependency-length growth is significantly different from the baseline dependency-length growth rate.[11] Using this method we find that dependency length in real sentences

[10] Note that this figure and several of the others are presented in full color in the electronic versions of this article, but in black and white in the print version; color versions of the figures are also available open access at http://muse.jhu.edu/resolve/98.

[11] We fit a mixed-effects regression model (Bresnan et al. 2007, Gelman & Hill 2007, Barr et al. 2013) with the following equation, with coefficients β representing fixed effects and coefficients S representing random effects by sentence.

$$\hat{y_i} = \beta_0 + S_0 + \beta_1 l_s^2 + (\beta_2 + S_2) r_i + \beta_3 r_i l_s^2 + \epsilon_i,$$

where $\hat{y_i}$ is the estimated total dependency length of data point i, β_0 is the intercept, l_s^2 is the squared length of sentence s in words, and r_i is an indicator variable, with value 1 if data point i is a random linearization and 0

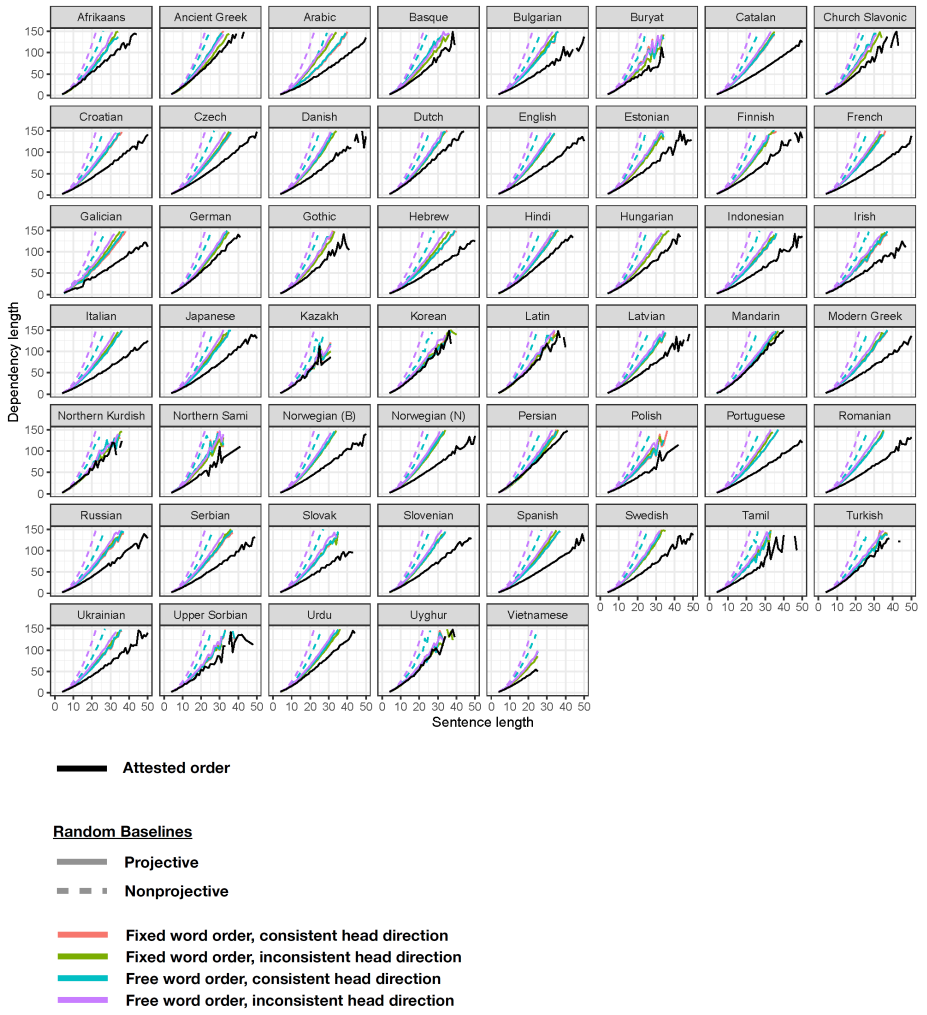


FIGURE 1. Dependency length as a function of sentence length for fifty-three languages. The *x*-axis is sentence length and the *y*-axis is the mean of total dependency length for all sentences of that length. The black line represents true dependency length. The colored lines represent random baselines. Solid colored lines represent projective baselines; dashed colored lines represent nonprojective baselines. Some baselines are not visible because they are covered by others.

grows more slowly than dependency length in random baselines for all baselines and languages at $p < 0.001$.

As a second, nonparametric statistical test, we used a sign test. For each real sentence and each baseline, we compared the dependency length in the real sentence to the average dependency length in the baseline reorderings. We assigned the real sentence a score of 1 if its dependency length was less than the average baseline dependency

if it is an observed linearization. We use l_s^2 rather than l_s following Futrell, Mahowald, & Gibson 2015, which found a better fit to the data with a squared predictor rather than a linear predictor. For significance testing comparing the real dependencies and random baselines, we performed a likelihood ratio test comparing models with and without β_3.

length, and 0 otherwise. A corpus is thus reduced to a sequence of 1s and 0s, where 1 indicates shorter dependency length than the random baseline; we can characterize a language with the proportion of sentences that have shorter dependency length than the random baseline. We then derive a sign-test *p*-value under the null hypothesis that the sequence of 1s and 0s was generated by random fair coin flips, following a binomial distribution. We consider a result significant if its *p*-value is less than 0.05.

The proportions of sentences with shorter dependencies than the baselines, and the results of the sign test, are shown in Figure 2. In the large majority of cases (53/53 languages at best, for the nonprojective free baseline, and 49/53 at worst, for the fixed random baseline), we find a significant DLM effect.[12]

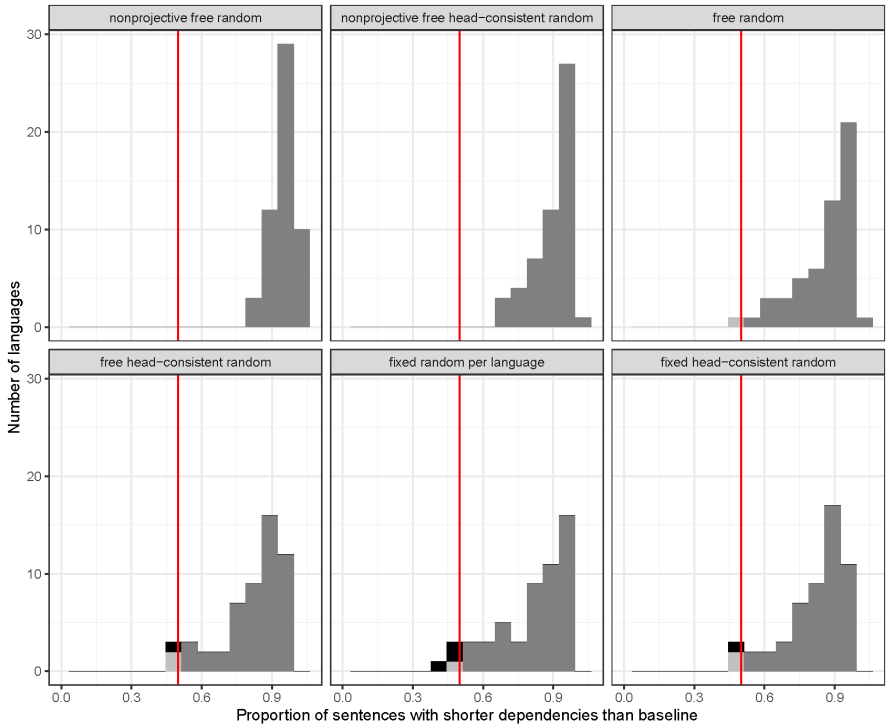


FIGURE 2. Histograms of languages by proportion of sentences with shorter dependencies than average baseline reorderings. One histogram per baseline. The vertical line indicates 50% on the *x*-axis. Darker gray blocks indicate languages where the significant majority of sentences have shorter dependencies than the baseline. Light gray indicates languages where the proportion of such sentences is not significantly different from one half. Black indicates languages where the significant majority of sentences in the corpus have longer dependency length than the baseline. The black exceptions are corpora of Uyghur (proportion of optimized sentences = 0.46), Latin (0.45), and Northern Kurdish (0.39) when compared to the fixed random baseline, and Korean when compared to the fixed (0.46) and free (0.47) head-consistent random baselines.

4.3. DISCUSSION. It is worth pausing to consider the meaning of Fig. 1. We can see a grammar and a set of usage preferences as defining a trajectory through the space

[12] Languages where dependency length is not significantly shorter than baselines are listed here: compared to the fixed head-consistent random baseline: Kazakh, Korean, Northern Kurdish; compared to the fixed random baseline: Ancient Greek, Latin, Northern Kurdish, Uyghur; compared to the free head-consistent random baseline: Kazakh, Korean, Northern Kurdish; compared to the free random baseline: Northern Kurdish.

shown in each facet of the figure. Some grammars and usage preferences will have steep trajectories, and others will have low trajectories. What Fig. 1 shows is that real languages tend to have grammars and usage preferences such that the trajectory is low, corresponding to low dependency length. If we think of each trajectory as a language, then we see DLM as a force pulling the languages downward.

We have shown that dependency length in natural language corpora is lower than would be expected given only constraints for projectivity, fixedness of word order, and consistency in head direction. These independent principles do not suffice to explain the low dependency length of natural language usage. The simplest explanation is potentially that DLM is in fact the causal factor behind all of these phenomena, though as we note below, our study does not exclude all other accounts.

The narrowest interpretation of our result is the following: whatever constraints exist on word order across languages, they have the effect of reducing dependency length beyond what would be expected from projectivity, consistency in head direction, and fixedness in word order. The simplest constraint that would accomplish this would be a direct constraint on dependency length itself. But it is also possible that the observed DLM effect emerges as a side-effect of some other, not-yet-explored factors. The way forward in future work is to implement more and more of these proposed other constraints on languages as baselines, and determine whether they can account for the empirical distribution of word orders.

In this connection, we note that the baselines we have presented here do not exhaust the possible independent constraints on language. Other constraints could include a complexity bound on grammars. There may also be other functionally motivated constraints on language, such as a requirement to allow robust information transmission in the presence of noise (Gibson et al. 2013, Futrell, Hickey, et al. 2015), or a requirement to maintain uniform information density (Fenk & Fenk 1980, Levy & Jaeger 2007, Jaeger 2010, Jaeger & Tily 2011; cf. Ferrer-i-Cancho et al. 2013). It is possible that random reorderings under some of these other constraints will produce dependency lengths comparable to natural language, but so far no combination of independently motivated constraints has sufficed to explain the shortness of dependency length in real utterances.

5. GRAMMAR AND USAGE. The studies in the previous section showed that dependency length in real sentences is lower than we would expect in baselines based on independently motivated constraints. These baselines show that dependency length is minimized beyond what we would expect if grammars and/or usage preferences were shaped by projectivity, consistent head direction, and fixedness in word order alone. However, they do not tell us whether this minimization happens in grammar, in usage preferences, or in both. The goal of this section is to disentangle these factors.

Figure 3 shows how both grammar and usage can result in an observed DLM preference in word order. Among all of the logically possible word orders for a tree expressing a particular meaning, the grammar of a language selects a set (or a probability distribution) of permitted orders. Then from that set, the language user selects one order to use. These two selections are two places where DLM can have an effect. For example, if the grammar permits only harmonic word orders, then the average utterance will come out with lower dependency length when compared with a grammar that enforces antiharmonic word orders (different branching directions for all dependency types). On the usage side, the grammar may permit either harmonic or antiharmonic orders, and the language user chooses the harmonic ones.

In terms of causal attribution, grammar and usage can never be separated with total certainty. Nevertheless, it is possible to use corpora to estimate the grammar—defined

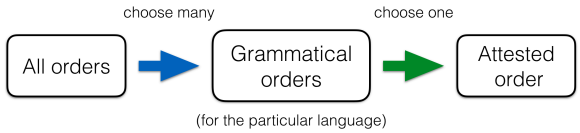


FIGURE 3. Schematic for how grammar and usage relate to linearizations. Grammar selects a set of permitted linearizations from the logically possible ones; usage selects one linearization from the grammatically permitted ones.

as the space of how trees can be linearized in general—and determine how that space relates to the observed linearization of any particular tree.

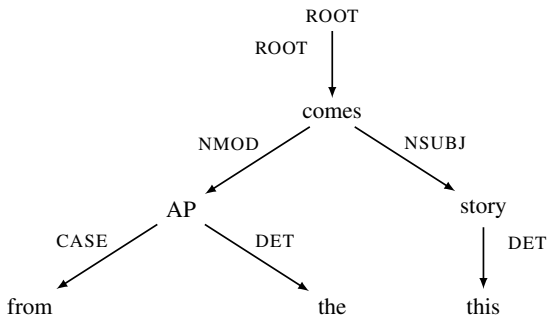
The goal of this section is to use probabilistic models of word order derived from corpora to argue that dependency locality affects both grammar and usage preferences. The logic is as follows. We take observed dependency trees and compare their dependency length to random reorderings according to the probabilistic model of the grammar. This work is essentially an attempt to automate the approach of Rajkumar et al. (2016), who compare dependency length in real utterances to dependency length in alternative grammatical utterances generated by hand. If the observed sentences have shorter dependency length than random grammatically possible reorderings, then this is evidence that language users are choosing particular utterances to minimize dependency length. Also, if the distribution of grammatical reorderings has lower dependency length than the random baselines from §4, then that is evidence that the grammar itself is affected by DLM.

Therefore, we are looking for two results: (i) whether observed sentences have shorter dependency length than the random grammatical reorderings (which shows DLM in usage), and (ii) whether the random grammatical reorderings have shorter dependency length than the independently motivated baselines (which shows DLM in grammar).

In what follows, we first describe the probabilistic models of word order we use (§5.1), and then present the results of comparing dependency length in real sentences and random grammatical reorderings (§5.2). Using this method we find evidence for DLM in both grammar and usage preferences.

5.1. LINEARIZATION MODELS. We define a LINEARIZATION MODEL as a stochastic function that takes as input an unordered dependency tree representation of a sentence and outputs an ordering of the words in the sentence. It is a model of the first step in Fig. 3. For example, 19 shows an unordered dependency tree: there exist head–dependent relations, but no notion of a word coming before or after any other word. Possible linearizations of this tree include (a) *This story comes from the AP* and (b) *From the AP comes this story*. The latter order is the attested order in the UD English corpus, but the former order will come out as much more likely under our ordering models.

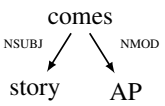
(19)



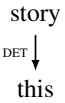
We focus on linearization models that can learn how to order dependency trees based on corpus data. In this work we make use of linearization models developed in the natural language processing (NLP) literature (Futrell & Gibson 2015). These models, in common with all linearization models from the NLP literature, can represent only a subset of the linearization rules present in the actual grammar of a language such as English. However, we argue below that they are capable of capturing many of the important patterns across languages.

GENERATIVE LINEARIZATION MODELS. Here we describe linearization models that are based on generative dependency models. We provide a brief description; a full description, including an experimental evaluation of the fluency and accuracy of the model output, can be found in Futrell & Gibson 2015.

Our linearization models capture the order distribution of the immediate dependents of a head. We call a head and its immediate dependents a LOCAL SUBTREE. Given an unordered tree such as the one in 19, we linearize it as follows. Starting from the first node under ROOT, the word *comes*, we consider its local subtree, shown in 20.

(20)


The model knows the observed corpus frequencies for the six possible orders of local subtrees consisting of a VERB head with NSUBJ and NMOD dependents, and it linearizes this local subtree randomly according to those statistics. Then the model moves to the head *story*, whose local subtree is shown in 21.

(21)


This local subtree is linearized by the same procedure, and so on recursively for the entire tree.

The resulting linearizations capture ordering constraints only within local subtrees. If in some language a head constrains the order of the children of one of its dependents, then these ordering constraints are not captured by this model. However, if the principle of endocentricity is to be believed, such constraints should be rare. The other major limitation of these ordering models is that they generate only word orders that are projective: they are incapable of generating discontinuous dependencies. We are willing to accept this limitation because most natural language word orders are projective.

There are a number of degrees of freedom in the specification of the model. For example, in the description above we said we collected ordering statistics for local subtrees defined in terms of a VERB head with NSUBJ and DOBJ dependents; it would also be possible to collect ordering statistics for local subtrees defined in terms of the part of speech of the dependent.

Futrell & Gibson 2015 evaluated the adequacy of a number of different model parameters in this setting. The parameters of variation in the model have to do with what information is taken into account when reordering trees. In the maximal model, trees are reordered based on parts of speech, dependency relation types, and the full set of dependents under the head. In the minimal model, trees are reordered based only on parts of speech, and each dependency is ordered without regard for the other dependencies under the current head. A large space of models can be defined at different levels of granularity between the minimal and maximal models, and it is possible to INTERPO-

LATE models: to create a model that combines the predictions of a more minimal and a more maximal model.

The previous work evaluated these model configurations on three points: first, whether they provide a good fit to the corpus data; second, whether they produce orders that humans find acceptable; and third, whether they produce orders that humans judge to have the same meaning as the original corpus sentence. The latter two criteria were evaluated only for the English models via Amazon Mechanical Turk. Under the best models, the previous work found average acceptability ratings of 3.8/5 (where the original sentences had average acceptabilities of 4.8/5), and the reordered sentences were judged to have the same meanings as the original sentences in 85% of cases.

Below, we use three model configurations of this kind to evaluate dependency locality in corpora. The first, most permissive configuration is a model that allows any linearization of a local subtree that has ever been observed in the corpus, defining local subtrees only in terms of dependency relation types, which we call PERMITTED orders. We also use the model that demonstrated the best fit to the corpus data across languages, which we call BEST FIT; this model is an interpolation of all possible model settings. Finally, we use the model that scored the highest for producing sentences with the same meaning as the original sentence in English, which we call SAME MEANING; this is the maximal model as described above.

NOTES ON INTERPRETATION. Before launching into the results, some discussion is in order on the specific linguistic interpretation of the random baselines defined by these linearization models.

The main constraint in the reordering models is that order is only computed relative to local subtrees, that is, the immediate dependents of a head. This is done in order to alleviate data sparsity in model estimation, but it puts limits on what ordering constraints can be represented by the model. In particular, it means that ordering constraints that involve heads and their grandchildren, or any other relationship going beyond direct head–dependent relationships and sibling relationships, are not represented. We also assume that linearizations are projective.

As such, the conservative interpretation of the results in this section is that we find DLM beyond what would be expected from only (i) projectivity and (ii) the ordering constraints among heads and immediate dependents. This limitation of our reordering model is equivalent to an assumption that language follows a context-free grammar. The results show conservatively that whatever constraints exist beyond what can be expressed in a context-free formalism, they serve to lower dependency length beyond what would be expected from projective dependency-local constraints alone. Nevertheless, because we believe a majority of word-order constraints can be represented in a context-free framework, we interpret the observed minimization of dependency length beyond the grammatical baselines as evidence for DLM in usage.

Our grammatical baselines reflect the average behavior at the level of the phrase. We believe this provides a useful estimate of grammatical constraints, but for those who do not take this as indicative of grammatical constraints, our results still show that people minimize dependency length in usage beyond what would be expected based on their average behavior at the level of single phrases.

5.2. RESULTS.

GENERATIVE LINEARIZATION MODELS. Figure 4 shows real dependency length compared to the random baselines for our custom linearization models. For this study, in order to avoid inaccurate ordering models, we exclude corpora with fewer than 1,500

sentences, meaning we exclude corpora of Buryat, Irish, Kazakh, Northern Kurdish, Tamil, Telugu, Upper Sorbian, and Uyghur beyond the previous exclusions.

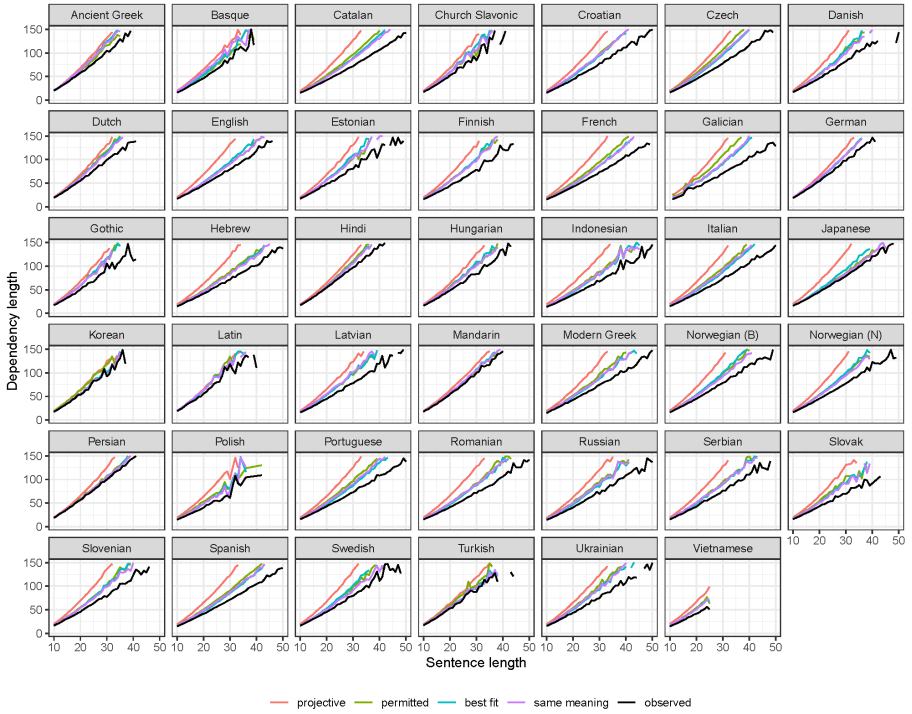


FIGURE 4. Mean dependency length as a function of sentence length, for real linearizations as compared with grammatical baselines.

The various grammatical baselines all produce linearizations with very similar dependency length. We do not attempt to draw any contrast among these baselines.

We see that the projective baseline has the longest dependency length, followed by the various grammatical baselines, followed by the observed dependency length. We analyzed the results statistically using the same regression methods described in §4.2. For all languages, the dependency-length growth rate for all of the baselines is greater than for the observed sentences at $p < 0.001$. Also, for all languages, the linearizations according to the simple baseline have a lower dependency-length growth rate than linearizations according to the projective baseline at $p < 0.001$ for all languages, suggesting that grammatical restrictions have the function of reducing dependency length.

The results show that grammatical orders are shorter than fully random orders, and that observed orders are shorter than grammatical orders. Thus, as a broad interpretation, we have evidence that both grammar and usage are affected by DLM: the observed DLM effect is explained partly by optimization of grammar and partly by optimization of usage. The most narrow interpretation is that people's expected ordering behavior at the level of the phrase minimizes dependency length, and that their behavior beyond what is described at the level of the phrase serves to further minimize dependency length.

5.3. DISCUSSION. We have demonstrated that dependency length is minimized in corpora beyond what would be expected from random grammatical reorderings of sentences as estimated by a probabilistic model. Furthermore, we showed that random grammati-

cal reorderings of sentences have lower dependency length than what would be expected from random baselines based on independently motivated constraints such as projectivity. While our results depend on the quality of our models of grammatical reorderings, we do not believe our model-construction process creates a bias for or against low dependency length.

The results give evidence for two hypotheses: first, that there are usage preferences in particular sentences that result in lower dependency length; and second, that the grammar of word order results in lower dependency length than would be expected from independently motivated constraints. These further provide evidence for the hypothesis that dependency locality has shaped both grammar and usage and can be used in a theory of word order.

In addition to our scientific point, we believe this work demonstrates a new and useful technique for corpus linguistics. Corpus linguistics has come under criticism for being able to study only usage, without being able to make claims about grammar (Newmeyer 1998). Our methodology provides a way to dissociate grammar and usage in corpus studies, by fitting probabilistic models of grammar based on corpora. Of course, no grammar induction method is perfect, and so we cannot draw conclusions about grammar with perfect certainty, but similar criticisms can also be leveled at the study of grammar through introspection (Hofmeister et al. 2013, Hofmeister et al. 2014, Hofmeister et al. 2015).

6. Variation in dependency length: head-finality. The studies above found evidence for a universal pressure toward short dependencies in grammar and usage. Every language tested showed a DLM effect compared to at least one baseline, and most languages showed a strong effect compared to all baselines. However, it is evident from Fig. 1 that languages differ in the extent of DLM. In this section we document some of the empirical variance in dependency length across languages and see how it correlates with other linguistic properties. We focus on the correlation between dependency length and head-finality in a language. We sketch some tentative hypotheses that could explain this variation, but we believe this phenomenon is fundamentally a puzzle for future theories of quantitative typology.

It is difficult to achieve a single corpus-based measure of dependency length that is comparable between corpora. The reason is that the range of possible dependency lengths for a sentence depends on the length of the sentence and on its dependency tree topology in a complex way (Ferrer-i-Cancho & Liu 2014, Esteban et al. 2016). For this reason, a simple summary statistic such as the mean dependency length per dependent (Liu 2008) can show effects across languages and corpora that do not reflect meaningful differences between languages, but rather only differences in sentence length.

In this work we dodge the question of how to develop a summary statistic for dependency length that is robust to variance in sentence length. Instead of developing such a measure, we present results comparing languages at fixed sentence lengths. Below, we compare dependency length in sentences of length ten, fifteen, and twenty words. Except where otherwise noted, the trends we present in this way are robust at other sentence lengths tested.

Below we show that dependency length in a language covaries with head-finality. For corpus evidence of covariance with word-order fixedness and morphological complexity, see Hawkins 1994:121–243 and Futrell 2017:118–22.

Figure 5 correlates dependency length with the proportion of head-final dependencies in a language, argued by Liu (2010) to be a reasonable reflection of the head-

direction typology of a language. As above, we exclude UD corpora with fewer than 1,500 sentences. Numerical values of the head-final proportions and average dependency length are given in Table 2. We find that more head-final languages have longer dependencies. Inspection of Fig. 1 confirms that many of the languages with especially long dependencies are predominantly head-final languages such as Japanese, Korean, and Turkish. Figure 5 also shows the correlation between the proportion of head-final dependencies and dependency length at sentence lengths ten, fifteen, and twenty words. Two correlation scores are reported: Pearson's *r*, which measures whether a linear relationship exists between head-finality and dependency length, and Spearman's rho, which measures whether any monotonic relationship exists between these variables. Both of these scores range in value from −1 (indicating a perfect inverse correlation) to 1 (indicating a perfect positive correlation), with 0 meaning no correlation is observed. The positive relationship between the proportion of head-final dependencies and dependency length is significant at all sentence lengths tested.

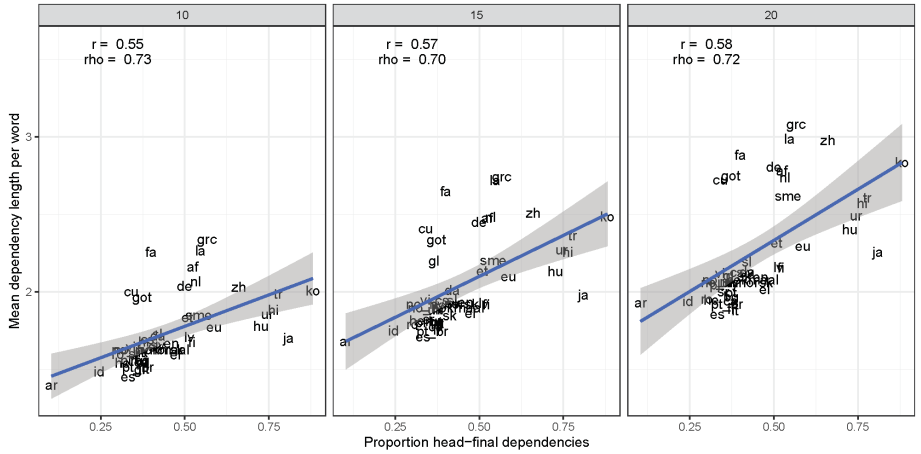


FIGURE 5. Languages by proportion of head-final dependencies (*x*-axis) and mean dependency length per word at sentence lengths 10, 15, and 20 (*y*-axis). Languages are represented by their two- or three-letter ISO codes. Pearson (*r*) and Spearman (rho) correlation coefficients are given.

In order to remove confounds involving the treatment of function words in the UD standard, we also calculated correlations of dependency length and the proportion of head-final dependencies after having removed all function words from the corpus.[13] The results of this analysis are shown in Figure 6, with numerical values given in Table 3.

This is not the first report of weaker DLM effects in head-final contexts. Rajkumar et al. (2016) also find a weaker locality preference among preverbal adjuncts in English as compared with other constructions.

6.1. POSSIBLE EXPLANATIONS. We believe the simplest explanation for this correlation has to do with another known bias in word-order preferences, the preference to place GIVEN material (material referring to discourse entities already discussed) before NEW material (which refers to new entities in the discourse: Prince 1981). The given-

[13] We removed all words having part of speech ADP, AUX, CCONJ, DET, PART, PRON, or SCONJ, or having a syntactic relation to their head of type AUX, CASE, CC, DET, EXPL, or MARK in the original UD annotation. We also ran the analysis of §4.2 having removed all function words in this way, finding the same pattern of results: see the appendix.

LANGUAGE	PROP. HF	DL@10	DL@15	DL@20	LANGUAGE	PROP. HF	DL@10	DL@15	DL@20
Korean	0.881	2.01	2.49	2.84	Norwegian (B)	0.401	1.63	1.90	2.08
Japanese	0.809	1.70	1.98	2.26	Persian	0.401	2.26	2.65	2.88
Turkish	0.778	1.99	2.36	2.61	Norwegian (N)	0.390	1.63	1.92	2.06
Hindi	0.763	1.88	2.26	2.57	Czech	0.389	1.69	1.94	2.13
Urdu	0.745	1.86	2.27	2.49	Italian	0.384	1.50	1.80	1.88
Hungarian	0.726	1.78	2.13	2.40	Croatian	0.380	1.68	1.89	2.06
Mandarin	0.661	2.03	2.51	2.98	French	0.374	1.51	1.75	1.89
Basque	0.587	1.77	2.10	2.29	Portuguese	0.373	1.55	1.81	2.00
Ancient Greek	0.566	2.34	2.74	3.08	Bulgarian	0.372	1.56	1.81	1.97
Latin	0.547	2.27	2.72	2.99	Gothic	0.372	1.97	2.34	2.75
Northern Sami	0.542	1.85	2.20	2.62	Catalan	0.371	1.55	1.78	1.94
Dutch	0.533	2.07	2.48	2.74	Ukrainian	0.368	1.61	1.89	2.06
Afrikaans	0.524	2.16	2.48	2.78	Galician	0.365	1.50	2.20	2.10
Finnish	0.521	1.67	1.92	2.16	Russian	0.358	1.56	1.81	2.07
Latvian	0.513	1.71	1.93	2.16	Serbian	0.349	1.60	1.82	2.00
Estonian	0.508	1.84	2.13	2.32	Church Slav.	0.341	2.00	2.41	2.72
German	0.500	2.04	2.45	2.81	Vietnamese	0.339	1.65	1.95	2.12
Modern Greek	0.472	1.59	1.86	2.02	Spanish	0.332	1.45	1.71	1.86
English	0.460	1.67	1.93	2.10	Polish	0.325	1.56	1.80	2.05
Danish	0.420	1.72	2.01	2.13	Hebrew	0.314	1.54	1.81	1.95
Swedish	0.420	1.66	1.93	2.13	Romanian	0.301	1.60	1.79	1.95
Slovenian	0.419	1.73	1.95	2.19	Indonesian	0.244	1.48	1.75	1.94
Slovak	0.412	1.65	1.85	2.10	Arabic	0.103	1.40	1.68	1.93

TABLE 2. Proportion of head-final dependencies (prop. HF) and mean dependency length per word for given sentence lengths (DL@10 = mean dependency length at sentence length 10, etc.).

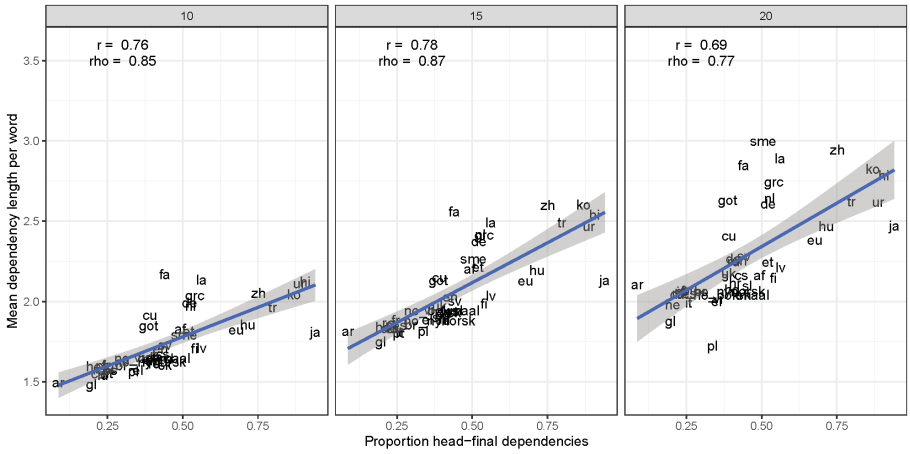


FIGURE 6. Languages by proportion of head-final dependencies (*x*-axis) and mean dependency length per word at sentence lengths 10, 15, and 20 (*y*-axis), FOR CONTENT WORDS ONLY. Pearson and Spearman correlation coefficients are given.

before-new bias is independent of DLM or any kind of locality principle, as far as is currently known. We discuss it further in §7.2. For now we note that if given material is usually expressed with shorter constituents and new material is usually expressed with longer constituents, then we would predict a universal short-before-long preference independently of DLM. This prediction was first made by Hawkins (1994:237–42). This independent bias would counteract the effects of DLM in head-final contexts, where DLM pushes for long-before-short orders. The overall result would be a weaker DLM

LANGUAGE	PROP. HF	DL@10	DL@15	DL@20	LANGUAGE	PROP. HF	DL@10	DL@15	DL@20
Japanese	0.941	1.81	2.13	2.47	Croatian	0.411	1.63	1.89	2.12
Hindi	0.906	2.13	2.54	2.79	Danish	0.407	1.66	1.93	2.27
Urdu	0.887	2.11	2.47	2.62	Russian	0.406	1.62	1.89	2.25
Korean	0.870	2.04	2.60	2.83	Bulgarian	0.401	1.63	1.92	2.08
Turkish	0.799	1.97	2.49	2.62	Norwegian (B)	0.399	1.65	1.95	2.04
Mandarin	0.752	2.05	2.60	2.95	Norwegian (N)	0.392	1.63	1.88	2.07
Hungarian	0.715	1.86	2.20	2.47	Church Slavonic	0.391	1.92	2.15	2.41
Basque	0.677	1.83	2.13	2.38	Ukrainian	0.390	1.65	1.97	2.18
Latvian	0.563	1.71	2.04	2.21	Gothic	0.387	1.85	2.13	2.63
Latin	0.562	2.14	2.49	2.89	Serbian	0.387	1.64	1.90	2.16
Ancient Greek	0.540	2.04	2.41	2.75	Vietnamese	0.355	1.66	1.95	2.02
Finnish	0.540	1.71	1.99	2.15	Modern Greek	0.351	1.58	1.89	2.00
Dutch	0.526	1.98	2.42	2.64	Polish	0.337	1.57	1.82	1.72
German	0.523	1.99	2.38	2.61	Spanish	0.259	1.57	1.85	2.05
Estonian	0.520	1.82	2.22	2.25	Italian	0.257	1.56	1.81	1.99
Northern Sami	0.505	1.79	2.27	3.00	Portuguese	0.255	1.57	1.81	2.07
Afrikaans	0.492	1.83	2.21	2.16	French	0.246	1.61	1.89	2.07
Slovenian	0.452	1.64	1.95	2.10	Indonesian	0.234	1.54	1.83	2.06
Swedish	0.442	1.73	2.01	2.28	Romanian	0.222	1.58	1.87	2.05
Persian	0.441	2.17	2.56	2.85	Catalan	0.219	1.55	1.83	2.05
Slovak	0.441	1.61	1.93	2.04	Hebrew	0.204	1.60	1.84	1.98
Czech	0.433	1.67	1.94	2.17	Galician	0.197	1.49	1.75	1.88
English	0.426	1.71	2.03	2.26	Arabic	0.087	1.50	1.82	2.11

TABLE 3. Proportion of head-final dependencies (prop. HF) and mean dependency length per word for given sentence lengths (DL@10 = mean dependency length at sentence length 10, etc.), with all function words removed.

effect for head-final dependencies. Evidence that elements referring to discourse-given referents are indeed associated with less DLM pressure in Mandarin is given by Xu and Liu (2015).

Recent literature in quantitative linguistics has suggested a cognitive motivation for longer dependencies in head-final contexts. The explanation is that placing more material before a head makes the head more predictable and thus easier to process when the comprehender reaches it (Ferrer-i-Cancho 2017). This benefit of delaying the verb counteracts the difficulty induced by the long dependencies. This explanation mirrors the fact that linguistic processing of sentence-final verbs does seem to get easier when more material appears before the verb, in what are known as antilocality effects (Konieczny 2000, Husain et al. 2014). For some theoretical issues with this explanation, see Levy 2005:79 and Futrell 2019.

Another explanation could have to do with morphology. Head-final languages typically have richer morphology than head-initial languages (Dryer 2002), perhaps because head-peripheral orders require case marking for robust information transmission in the presence of noise (Gibson et al. 2013, Futrell, Hickey, et al. 2015). Morphology (case and/or agreement) provides informative cues about what the head of each marked word is. If we think that dependency locality effects are in part driven by inaccuracy in memory retrieval during parsing (as argued in Vasishth et al. 2017), then such morphology would alleviate dependency locality effects. Indeed, Ros et al. (2015) find weaker DLM preferences in morphologically rich languages.

7. CONCLUSION. We have given extensive quantitative evidence for dependency locality as a shaper of word order across languages in both grammar and usage. We have shown that dependency-length minimization is a macroscale property of syntactic trees across many languages, in that dependency length is shorter than random baselines in-

corporating both independently motivated constraints and language-specific grammatical constraints. Also, we documented substantial variance in dependency length across languages, which appears to be correlated with head-finality. We consider the explanation of this variance an open question.

7.1. PROSPECTS FOR DEPENDENCY LOCALITY AND RELATED THEORIES. We believe this work, along with other similar evidence, suffices to establish dependency locality as a principle of natural language word order. The question then becomes: what next? We believe the most promising direction for future work on the efficiency hypothesis is to find phenomena that cannot be explained by dependency locality and determine either what other factors explain those phenomena or whether they can be explained by some generalization of dependency locality.

One promising extension of dependency locality theory comes from taking seriously the results from psycholinguistics indicating that the bulk of language processing load comes from the degree to which linguistic elements such as words are unexpected in context. Surprisal theory (Hale 2001, Levy 2008, Smith & Levy 2013) formalizes this idea and claims that ALL processing difficulty results from the extent to which elements are unexpected in context given the comprehender's knowledge of the usage distribution. Surprisal theory in its usual form cannot account for locality effects (Levy 2013); however, a recent extension of the theory does predict dependency locality effects in a generalized form. The theory of LOSSY-CONTEXT SURPRISAL holds that processing difficulty results from the extent to which linguistic elements are unpredictable given a NOISY MEMORY REPRESENTATION of context (Futrell & Levy 2017, Futrell 2019); this theory gives rise to a principle called INFORMATION LOCALITY, which is that processing difficulty results whenever pairs of linguistic elements that predict each other are far from each other. Head–dependent pairs fall into this category (Futrell et al. 2019); thus dependency locality can be seen as a special case of information locality.

Information locality subsumes the predictions of dependency locality and also makes a number of fine-grained predictions in domains where dependency locality makes none. For example, information locality is an accurate predictor of adjective ordering in noun phrases with multiple adjectives (Futrell 2019); dependency locality makes no predictions about such adjectives (assuming a syntactic structure where all of the adjectives are codependent on the head noun). For alternative information-theoretic explanations of adjective order constraints, see Dyer 2017, Hahn et al. 2018, and Scontras et al. 2019. Information locality also predicts more generally that, among words in dependencies, those word pairs that predict each other more will be under even stronger pressure to be close to each other, which prediction is borne out in corpora (Futrell 2019).

While information locality successfully extends dependency locality, certain word-order patterns, such as those documented in §6, remain unexplained. Another unexplained pattern is that languages with more morphological complexity seem to show weaker dependency locality effects in on-line processing (Ros et al. 2015) and in word order (Gulordava & Merlo 2015). These effects seem to have an intuitive explanation—a language with more morphological marking may be less reliant on adjacency to convey syntactic structure—but this explanation has yet to be formalized computationally so that it can be tested quantitatively in corpora.

7.2. BEYOND LOCALITY. There is another major class of word-order biases that do not admit an explanation in terms of any kind of locality theory. Here we discuss these phenomena and how they relate to our theory and findings. These are theories involving left-right order asymmetries, primarily preferences to put certain elements early; the theory of dependency locality is symmetrical in that it predicts no such left-right

asymmetries. We consider the explanation of these ordering preferences to be open problems.

A prominent subset of these other biases have to do with tendencies for certain items to be placed earlier in a sentence. The items that are biased to appear early are:

(i) GIVEN items, which refer to discourse referents already established, as opposed to new items (Prince 1981) (but see Derbyshire 1979 for evidence that Hixkaryana has a new-before-given bias).

(ii) Items that are ANIMATE and/or DEFINITE, essentially following the 'animacy hierarchy' (Silverstein 1976, Kiparsky 2008). Across constructions, animate and definite items are placed earlier (McDonald et al. 1993), for example in the dative and genitive alternations in English (Bresnan et al. 2007, Shih et al. 2015).

As discussed in §6, these biases could provide an explanation for some of the observed crosslinguistic variation in dependency length. In particular, the bias for long-before-short orders in head-final contexts would be weakened by the given-before-new bias, under the assumption that constituents referring to given items are typically shorter than constituents referring to new items.

The most common explanation for these biases is that they reflect an easy-first strategy in utterance production and planning. That is, when people are formulating and producing utterances, they produce the things that are easy to produce earlier, as part of a general greedy production strategy (Bock 1982). Another way of describing this theory is to posit that certain noun phrases, such as the given, definite, and animate ones, are more ACCESSIBLE during processing (Ariel 1990). These theories are bolstered by behavioral evidence: animate nouns appear to be faster to retrieve from memory in many circumstances (Popp & Serra 2016). Given nouns can be seen as subject to a kind of priming effect, where they become easier to produce because a similar expression was produced recently. The intuition that these word-order biases can arise from an easy-first production strategy has been formalized in a neural network model by Chang (2009).

It is an open question whether and how these biases can be derived from a general theory of processing cost, in a way that may predict interesting interactions with locality pressures. Along these lines, Hawkins (2004) proposes a non-locality-based processing pressure that has a similar flavor: MAXIMIZE ONLINE PROCESSING (MaOP). MaOP holds that the human language processor prefers to assign all syntactic and semantic properties to a form as soon as possible, and thus words with more explicit marking will appear earlier. This principle has been used to explain why, for example, fillers usually precede gaps in filler-gap constructions. We believe it will be fruitful to attempt to formalize this concept so that it can be applied in quantitative corpus studies.

7.3. CONCLUSION. We have presented extensive studies using corpora and probabilistic grammars to argue that natural language word order is shaped by a simple pressure for dependency locality. We believe this macroscale, quantitative approach to linguistic analysis complements traditional microscale and qualitative formal methods: the large-scale studies can reveal the gross constraints that affect languages, and the microscale studies can reveal how these constraints are satisfied and instantiated in myriad different structural configurations.

We believe our results show that the efficiency hypothesis is a promising means for explaining language universals. Natural language has many properties that are unlike any other known communication system; it stands to reason that these properties arise because natural language must operate under the unique constraints imposed by the human brain in both learning and on-line processing.

Appendix: Dependency length under other transformations of UD

Here we present the results from §4 under different transformations of the dependency corpora. The results in the previous section had undergone the transformations and filters defined in §3.3. Here we present two results. First, we present results from using the UD dependency trees in their original content-head form, where content words are always heads of function words. Second, we present results from using the UD dependency trees while removing all function words, thus removing all potential bias due to the parsing of function words and establishing that a DLM effect exists solely between content words.

A1. With original UD parses. Figure A1 shows the same result as Fig. 1, but using the original content-head dependencies from UD. Corpora are transformed only to remove punctuation and collapse 'flat' dependencies such as multiword expressions and foreign phrases. Figure 2 is repeated in Figure A2 for content-head dependencies.

The results are largely similar to the results using function-word-head dependencies. The statistical significance of the main DLM result is also maintained: the dependency-length growth rate in all languages is significantly slower for observed sentences than for all random baselines.

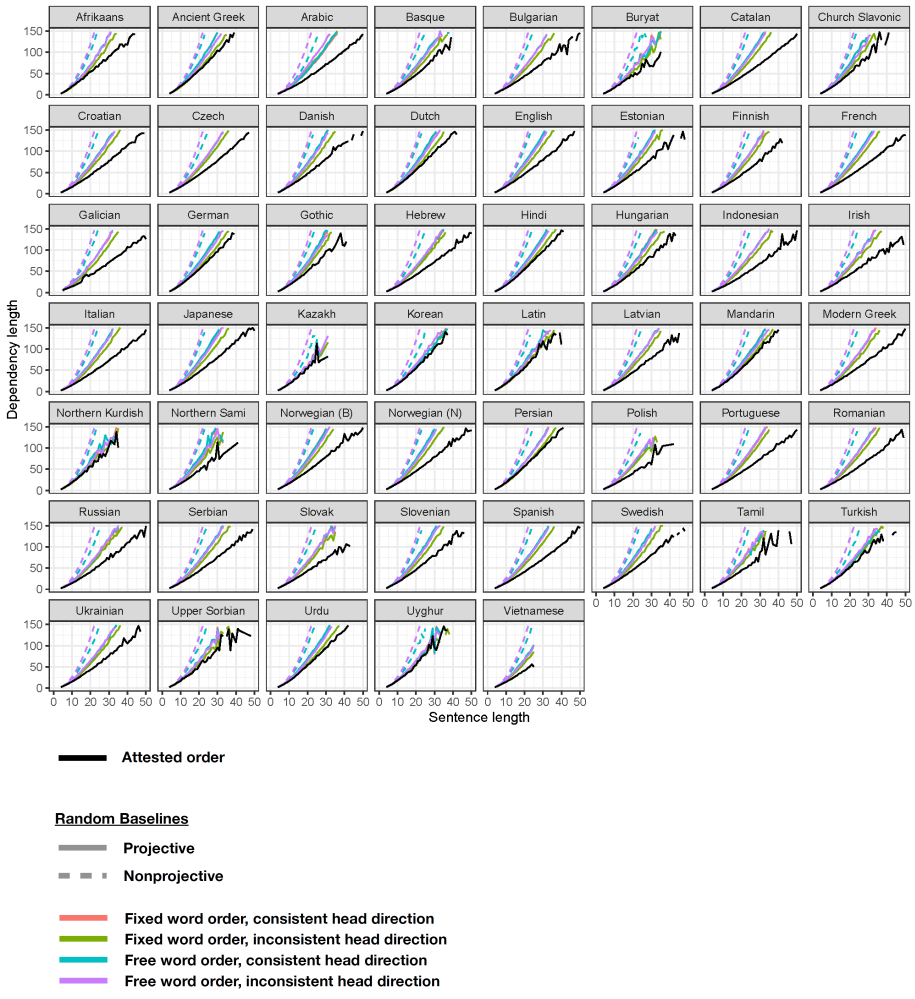


Figure A1. Mean dependency length as a function of sentence length for fifty-three languages, according to original UD parses. The black line represents true dependency length. The colored lines represent random baselines. Solid colored lines represent projective baselines; dashed colored lines represent nonprojective baselines. Some baselines are not visible because they are covered by others.

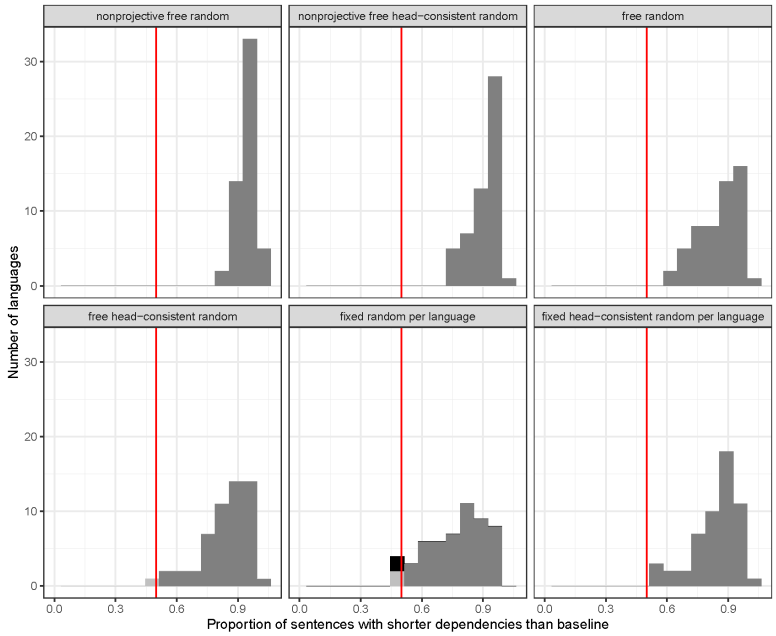


FIGURE A2. Histograms of languages by proportion of sentences with shorter dependencies than average baseline reorderings, ACCORDING TO ORIGINAL UD PARSES. Shading means the same as in Fig. 2. The black exceptions are corpora of Ancient Greek (proportion of optimized sentences = 0.49) and Latin (0.45) when compared to the fixed random baseline.

It may seem surprising that we still find a significant DLM effect in the original content-head trees, because content-head annotation distorts many of the dependencies whose actual order in natural language has been argued to support DLM. For example, the content-head dependency length of an example such as in 16 above could be improved by introducing a disharmonic order, putting the adposition *on* after the noun *internet*. Nevertheless, we find that the DLM result is robust to content-head annotation: we find the significant DLM effect even for such dependency trees.

We believe that our result is robust to content-head annotation because it primarily reflects a DLM effect among large phrases. When large constituents, such as multiple locative adjuncts modifying a verb, are ordered contrary to DLM, then there is the potential to dramatically increase the sum dependency length of a sentence. In contrast, if the noun, determiner, and preposition inside a three-word PP are ordered contrary to DLM, then this will have a smaller numerical effect on the dependency length of the sentence. In fact, the available corpus evidence suggests that DLM effects in short spans are somewhat variable and complex (Gulordava et al. 2015), in contrast to the simple effect demonstrated in this article. If further studies bear out the idea that DLM primarily affects the order of large constituents, with weaker effects in short spans, then that would suggest that DLM is a relatively weak (but pervasive) force on word order.

It remains an open question whether DLM effects are truly stronger in some sense over large spans as opposed to short spans; this question is related to whether the cognitive cost associated with long dependencies scales linearly, superlinearly, or sublinearly with respect to dependency length.

There is one major difference between the content-head results and the function-word-head results: the fixed-word-order inconsistent-head-direction baseline (the green line) consistently outperforms the other baselines across languages in terms of dependency length. In fact, when we closely compare Fig. A1 and Fig. 1, we can see that this change is due to the other baselines becoming worse under content-head dependencies, while the green baseline remains about the same.

We believe the best explanation for this pattern is that DLM favors inconsistency in head direction when trees have a large number of dependents per head, as discussed in §2.3. Content-head trees typically have more dependents per head than function-word-head trees, so their dependency length is reduced when inconsistency in head direction is allowed in this random baseline. Furthermore, having fixed word order means that if a language has this desirable inconsistency in one PP, it is likely to have it in all PPs (which are built out of the same dependency relation types), thus decreasing the dependency length in this baseline.

A2. REMOVING FUNCTION WORDS. Here we show the same results as in Fig. 1, but with all function words removed from the corpora. Function words are defined for these purposes as words with part-of-speech tags

ADP, AUX, CCONJ, DET, PART, PRON, and SCONJ, or words whose dependency relation type to their head is AUX, CASE, CC, DET, EXPL, or MARK.

This study allows us to answer three questions. First: is the significant DLM effect in previous results an artifact of the way function words are parsed in UD? Second: is there a significant DLM effect across languages when we consider only content words? Third: can the variance in the strength of the DLM effect between languages be explained in terms of the relative abundance of function words?

The last question deserves some further explanation. If a language has many function words that typically appear close to their dependents, such as prepositions and determiners in English, then that language might appear to have a stronger DLM effect than a language such as Russian where function words are rarer, simply because there exist many short dependencies in English with no analogue in Russian. By looking at the strength of the DLM effect ignoring function words, we can control for this potential confound.

Figure A3 shows the comparison of observed dependency length with the random baselines while ignoring function words. The DLM effect remains in all languages and is statistically significant in all languages, so a DLM effect can be established purely among content words, and the previous results were not dependent on the particular parsing decisions for function words.

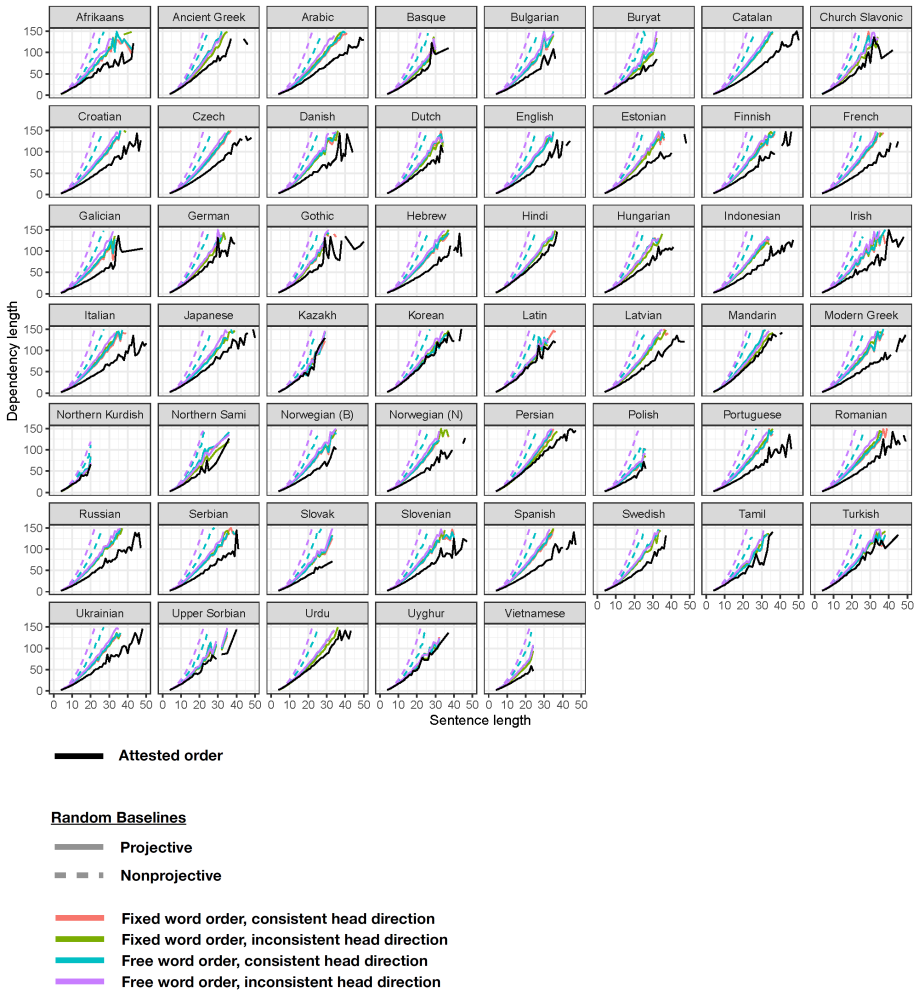


FIGURE A3. Mean dependency length as a function of sentence length for fifty-three languages, IGNORING ALL FUNCTION WORDS. The black line represents true dependency length. The colored lines represent random baselines. Solid colored lines represent projective baselines; dashed colored lines represent nonprojective baselines. Some baselines are not visible because they are covered by others.

A great deal of variance in the strength of DLM still exists between languages. However, some of the salient differences are made smaller. For example, observe the dependency-length characteristics for Japanese and Korean. Japanese and Korean are languages with very similar syntax, but they end up with wildly different dependency-length characteristics in Fig. 1. We believe the reason for this difference lies in the way that function words are parsed in the two languages. In the Japanese corpus, all morphemes are treated as separate tokens, so morphologically complex 'words' are split into multiple tokens with many short dependencies. Thus the Japanese corpus shows short dependency length. In Korean, by contrast, morphologically complex words are considered units, so the corresponding short dependencies do not exist. Thus the Korean corpus shows long dependency length. With function words removed, the gap between Japanese and Korean shrinks dramatically, indicating that the presence of function words was driving a large portion of the difference between those languages.

REFERENCES

ABNEY, STEVEN PAUL. 1987. *The English noun phrase in its sentential aspect*. Cambridge, MA: MIT dissertation.

ADGER, DAVID. 2003. *Core syntax: A minimalist approach*. Oxford: Oxford University Press.

ALEXIADOU, ARTEMIS; LILIANE HAEGEMAN; and MELITA STAVROU. 2007. *Noun phrase in the generative perspective*. Berlin: Mouton de Gruyter.

ANTTILA, ARTO; MATTHEW ADAMS; and MICHAEL SPERIOSU. 2010. The role of prosody in the English dative alternation. *Language and Cognitive Processes* 25.946–81. DOI: 10.1080/01690960903525481.

ARIEL, MIRA. 1990. *Accessing noun-phrase antecedents*. London: Routledge.

BAKER, MARK C. 2001. *The atoms of language: The mind's hidden rules of grammar*. New York: Basic Books.

BARR, DALE J.; ROGER LEVY; CHRISTOPH SCHEEPERS; and HARRY J. TILY. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68.255–78. DOI: 10.1016/j.jml.2012.11.001.

BARTEK, BRIAN; RICHARD L. LEWIS; SHRAVAN VASISHTH; and MASON R. SMITH. 2011. In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37.1178–98. DOI: 10.1037/a0024194.

BEHAGHEL, OTTO. 1909. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen* 25.110–42. DOI: 10.1515/9783110242652.110.

BEHAGHEL, OTTO. 1930. Zur Wortstellung des Deutschen. *Curme volume of linguistic studies* (*Language* monograph 7), ed. by James Taft Hatfield, Werner Leopold, and A. J. Friedrich Zieglschmid, 29–33. Baltimore: Waverly. DOI: 10.2307/521983.

BENOR, SARAH, and ROGER LEVY. 2006. The chicken or the egg? A probabilistic analysis of English binomials. *Language* 82.233–78. DOI: 10.1353/lan.2006.0077.

BHATT, RAJESH, and ARAVIND K. JOSHI. 2004. Semilinearity is a syntactic invariant: A reply to Michaelis and Kracht 1997. *Linguistic Inquiry* 35.683–92. DOI: 10.1162/ling.2004.35.4.683.

BLOOMFIELD, LEONARD. 1933. *Language*. New York: Henry Holt.

BOCK, J. KATHRYN. 1982. Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review* 89.1–47. DOI: 10.1037/0033-295X.89.1.1.

BRESNAN, JOAN. 1982. *The mental representation of grammatical relations*. Cambridge, MA: MIT Press.

BRESNAN, JOAN; ANNA CUENI; TATIANA NIKITINA; and HARALD BAAYEN. 2007. Predicting the dative alternation. *Cognitive foundations of interpretation*, ed. by Gerlof Bouma and Joost Zwarts, 69–94. Amsterdam: Royal Netherlands Academy of Science.

BRUENING, BENJAMIN. 2009. Selectional asymmetries between CP and DP suggest that the DP hypothesis is wrong. *University of Pennsylvania Working Papers in Linguistics* 15.27–35. Online: https://repository.upenn.edu/pwpl/vol15/iss1/5.

CHANG, FRANKLIN. 2009. Learning to order words: A connectionist model of heavy NP shift and accessibility effects in Japanese and English. *Journal of Memory and Language* 61.374–97. DOI: 10.1016/j.jml.2009.07.006.

CHOMSKY, NOAM. 1959. On certain formal properties of grammars. *Information and Control* 2.137–67. DOI: 10.1016/S0019-9958(59)90362-6.

Chomsky, Noam. 1975. *The logical structure of linguistic theory*. Chicago: University of Chicago Press.
Chomsky, Noam. 1981. *Lectures on government and binding*. Dordrecht: Foris.
Chomsky, Noam. 1995. *The minimalist program*. (Current studies in linguistics 28.) Cambridge: Cambridge University Press.
Chomsky, Noam. 2000. Minimalist inquiries: The framework. *Step by step: Essays on minimalist syntax in honor of Howard Lasnik*, ed. by Roger Martin, David Michaels, and Juan Uriagereka, 89–155. Cambridge, MA: MIT Press.
Chomsky, Noam. 2004. Beyond explanatory adequacy. *Structures and beyond*, ed. by Adriana Belletti, 104–31. Oxford: Oxford University Press.
Chomsky, Noam. 2005. Three factors in language design. *Linguistic Inquiry* 36.1–61. DOI: 10.1162/0024389052993655.
Chomsky, Noam. 2007. Approaching UG from below. *Interfaces + recursion = language?: Chomsky's minimalism and the view from syntax-semantics*, ed. by Uli Sauerland and Hans-Martin Gärtner, 1–29. Berlin: Mouton de Gruyter. DOI: 10.1515/9783110207552.1.
Chomsky, Noam, and Howard Lasnik. 1977. Filters and control. *Linguistic Inquiry* 8.425–504. Online: https://www.jstor.org/stable/4177996.
Chomsky, Noam, and Marcel P. Schützenberger. 1963. The algebraic theory of context free languages. *Computer programming and formal languages*, ed. by P. Braffort and D. Hirschberg, 118–61. Amsterdam: North-Holland.
Christiansen, Morten H., and Nick Chater. 2008. Language as shaped by the brain. *Behavioral and Brain Sciences* 31.489–509. DOI: 10.1017/S0140525X08004998.
Chung, Fan-Rong King. 1984. On optimal linear arrangements of trees. *Computers & Mathematics with Applications* 10.43–60. DOI: 10.1016/0898-1221(84)90085-3.
Corbett, Greville G.; Norman M. Fraser; and Scott McGlashan (eds.) 1993. *Heads in grammatical theory*. Cambridge: Cambridge University Press.
Courtin, Marine, and Chunxiao Yan. 2019. What can we learn from natural and artificial dependency trees. *Proceedings of the First Workshop on Quantitative Syntax*, 125–35. DOI: 10.18653/v1/W19-7915.
Cover, Thomas M., and J. A. Thomas. 2006. *Elements of information theory*. Hoboken, NJ: John Wiley & Sons.
Croft, William A.; Dawn Nordquist; Katherine Looney; and Michael Regan. 2017. Linguistic typology meets universal dependencies. *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories*, 63–75. Online: http://ceur-ws.org/Vol-1779/05croft.pdf.
Crutchfield, James P., and Karl Young. 1989. Inferring statistical complexity. *Physical Review Letters* 63.105–8. DOI: 10.1103/PhysRevLett.63.105.
Culbertson, Jennifer, and David Adger. 2014. Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences* 111.5842–47. DOI: 10.1073/pnas.1320525111.
Dębowski, Łukasz. 2011. Excess entropy in natural language: Present state and perspectives. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 21:037105. DOI: 10.1063/1.3630929.
Derbyshire, Desmond C. 1979. *Hixkaryána*. (*Lingua* descriptive series 1.) Amsterdam: North-Holland.
Dryer, Matthew S. 1992. The Greenbergian word order correlations. *Language* 68.81–138. DOI: 10.2307/416370.
Dryer, Matthew S. 2002. Case distinctions, rich verb agreement, and word order type (Comments on Hawkins' paper). *Theoretical Linguistics* 28.151–58. DOI: 10.1515/thli.2002.28.2.151.
Dyer, William E. 2017. *Minimizing integration cost: A general theory of constituent order*. Davis: University of California, Davis dissertation. Online: https://ucdavis.app.box.com/s/h4yk2plex1f4ls7d99zf0myr8ttxlkjv.
Esteban, Juan Luis; Ramon Ferrer-i-Cancho; and Carlos Gómez-Rodríguez. 2016. The scaling of the minimum sum of edge lengths in uniformly random trees. *Journal of Statistical Mechanics: Theory and Experiment* 16:063401. DOI: 10.1088/1742-5468/2016/06/063401.

FEDZECHKINA, MARYIA; BECKY CHU; and T. FLORIAN JAEGER. 2017. Human information processing shapes language change. *Psychological Science* 29.72–82. DOI: 10.1177/0956797617728726.

FEDZECHKINA, MARYIA; T. FLORIAN JAEGER; and ELISSA L. NEWPORT. 2012. Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences* 109.17897–902. DOI: 10.1073/pnas.1215776109.

FENK, AUGUST, and GERTRUD FENK. 1980. Konstanz im Kurzzeitgedächtnis—Konstanz im sprachlichen Informationsfluß. *Zeitschrift für experimentelle und angewandte Psychologie* 27.400–414.

FERRER-I-CANCHO, RAMON. 2004. Euclidean distance between syntactically linked words. *Physical Review E* 70:056135. DOI: 10.1103/PhysRevE.70.056135.

FERRER-I-CANCHO, RAMON. 2006. Why do syntactic links not cross? *Europhysics Letters* 76:1228. DOI: 10.1209/epl/i2006-10406-0.

FERRER-I-CANCHO, RAMON. 2016. Non-crossing dependencies: Least effort, not grammar. *Towards a theoretical framework for analyzing complex linguistic networks*, ed. by Alexander Mehler, Andy Lücking, Sven Banisch, Philippe Blanchard, and Barbara Job, 203–34. Berlin: Springer. DOI: 10.1007/978-3-662-47238-5_10.

FERRER-I-CANCHO, RAMON. 2017. The placement of the head that maximizes predictability: An information theoretic approach. *Glottometrics* 39.38–71.

FERRER-I-CANCHO, RAMON; ŁUKASZ DĘBOWSKI; and FERMÍN MOSCOSO DEL PRADO MARTÍN. 2013. Constant conditional entropy and related hypotheses. *Journal of Statistical Mechanics: Theory and Experiment* 2013:L07001. DOI: 10.1088/1742-5468/2013/07/L07001.

FERRER-I-CANCHO, RAMON, and CARLOS GÓMEZ-RODRÍGUEZ. 2016. Crossings as a side effect of dependency lengths. *Complexity* 21.320–28. DOI: 10.1002/cplx.21810.

FERRER-I-CANCHO, RAMON; CARLOS GÓMEZ-RODRÍGUEZ; and JUAN LUIS ESTEBAN. 2018. Are crossing dependencies really scarce? *Physica A: Statistical Mechanics and its Applications* 493.311–29. DOI: 10.1016/j.physa.2017.10.048.

FERRER-I-CANCHO, RAMON, and HAITAO LIU. 2014. The risks of mixing dependency lengths from sequences of different length. *Glottotheory* 5.143–55. DOI: 10.1515/glot-2014-0014.

FUTRELL, RICHARD. 2017. *Memory and locality in natural language*. Cambridge, MA: MIT dissertation. Online: http://hdl.handle.net/1721.1/114075.

FUTRELL, RICHARD. 2019. Information-theoretic locality properties of natural language. *Proceedings of the First Workshop on Quantitative Syntax*, 2–15. DOI: 10.18653/v1/W19-7902.

FUTRELL, RICHARD, and EDWARD GIBSON. 2015. Experiments with generative models for dependency tree linearization. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1978–83. DOI: 10.18653/v1/D15-1231.

FUTRELL, RICHARD; TINA HICKEY; ALDRIN LEE; EUNICE LIM; ELENA LUCHKINA; and EDWARD GIBSON. 2015. Cross-linguistic gestures reflect typological universals: A subject-initial, verb-final bias in speakers of diverse languages. *Cognition* 136.215–21. DOI: 10.1016/j.cognition.2014.11.022.

FUTRELL, RICHARD, and ROGER LEVY. 2017. Noisy-context surprisal as a human sentence processing cost model. *Proceedings of the 15th conference of the European Chapter of the Association for Computational Linguistics, vol. 1: Long papers*, 688–98. Online: https://www.aclweb.org/anthology/E17-1065.

FUTRELL, RICHARD; KYLE MAHOWALD; and EDWARD GIBSON. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences* 112.10336–341. DOI: 10.1073/pnas.1502134112.

FUTRELL, RICHARD; PENG QIAN; EDWARD GIBSON; EVELINA FEDORENKO; and IDAN BLANK. 2019. Syntactic dependencies correspond to word pairs with high mutual information. *Proceedings of the Fifth International Conference on Dependency Linguistics (DepLing 2019)*, 3–13. DOI: 10.18653/v1/W19-7703.

GELMAN, ANDREW, and JENNIFER HILL. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511790942.

Gerdes, Kim; Bruno Guillaume; Sylvain Kahane; and Guy Perrier. 2018. SUD or surface-syntactic universal dependencies: An annotation scheme near-isomorphic to UD. *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, 66–74. DOI: 10.18653/v1/W18-6008.

Gerdes, Kim; Bruno Guillaume; Sylvain Kahane; and Guy Perrier. 2019. Improving surface-syntactic universal dependencies (SUD): Surface-syntactic relations and deep syntactic features. *Proceedings of the 18th International Workshop on Treebanks & Linguistic Theory*, 126–32. DOI: 10.18653/v1/W19-7814.

Gibson, Edward. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition* 68.1–76. DOI: 10.1016/S0010-0277(98)00034-1.

Gibson, Edward. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain: Papers from the first Mind Articulation Project Symposium*, ed. by Alec Marantz, Yasushi Miyashita, and Wayne O'Neil, 95–126. Cambridge, MA: MIT Press.

Gibson, Edward; Richard Futrell; Steven T. Piantadosi; Isabelle Dautriche; Kyle Mahowald; Leon Bergen; and Roger Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences* 23.389–407. DOI: 10.1016/j.tics.2019.02.003.

Gibson, Edward; Steven T. Piantadosi; Kimberly Brink; Leon Bergen; Eunice Lim; and Rebecca Saxe. 2013. A noisy-channel account of crosslinguistic word-order variation. *Psychological Science* 24.1079–88. DOI: 10.1177/0956797612463705.

Gildea, Daniel, and T. Florian Jaeger. 2015. Human languages order information efficiently. arXiv:1510.02823 [cs.CL]. Online: http://arxiv.org/abs/1510.02823.

Gildea, Daniel, and David Temperley. 2007. Optimizing grammars for minimum dependency length. *Proceedings of the 45th annual meeting of the Association for Computational Linguistics*, 184–91. Online: http://www.aclweb.org/anthology/P07-1024.

Gildea, Daniel, and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science* 34.286–310. DOI: 10.1111/j.1551-6709.2009.01073.x.

Givón, Talmy. 1991. Markedness in grammar: Distributional, communicative and cognitive correlates of syntactic structure. *Studies in Language* 15.335–70. DOI: 10.1075/sl.15.2.05giv.

Gómez-Rodríguez, Carlos; Morten H. Christiansen; and Ramon Ferrer-i-Cancho. 2019. Memory limitations are hidden in grammar. arXiv:1908.06629 [cs.CL]. Online: https://arxiv.org/abs/1908.06629.

Greenberg, Joseph H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, ed. by Joseph H. Greenberg, 73–113. Cambridge, MA: MIT Press.

Grodner, Daniel, and Edward Gibson. 2005. Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science* 29.261–90. DOI: 10.1207/s15516709cog0000_7.

Gross, Thomas, and Timothy Osborne. 2009. Toward a practical dependency grammar theory of discontinuities. *SKY Journal of Linguistics* 22.43–90. Online: http://www.ling.helsinki.fi/sky/julkaisut/SKY2009/Gross_Osborne_NETTI.pdf.

Gulordava, Kristina, and Paola Merlo. 2015. Diachronic trends in word order freedom and dependency length in dependency-annotated corpora of Latin and Ancient Greek. *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, 121–30. Online: https://www.aclweb.org/anthology/W15-2115.

Gulordava, Kristina; Paola Merlo; and Benoit Crabbé. 2015. Dependency length minimisation effects in short spans: A large-scale analysis of adjective placement in complex noun phrases. *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, vol. 2: Short papers*, 477–82. DOI: 10.3115/v1/P15-2078.

Hahn, Michael; Judith Degen; Noah Goodman; Daniel Jurafsky; and Richard Futrell. 2018. An information-theoretic explanation of adjective ordering preferences. *Proceedings of the 40th annual meeting of the Cognitive Science Society (CogSci 2018)*, 1766–71. Online: https://cogsci.mindmodeling.org/2018/papers/0339/0339.pdf.

Hale, John T. 2001. A probabilistic Earley parser as a psycholinguistic model. *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics*. Online: https://www.aclweb.org/anthology/N01-1021.

HALE, KENNETH. 1983. Warlpiri and the grammar of non-configurational languages. *Natural Language and Linguistic Theory* 1.5–47. DOI: 10.1007/BF00210374.
HAVELKA, JIŘÍ. 2007. Beyond projectivity: Multilingual evaluation of constraints and measures on non-projective structures. *Proceedings of the 45th annual meeting of the Association of Computational Linguistics*, 608–15. Online: https://www.aclweb.org/anthology/P07-1077.
HAWKINS, JOHN A. 1990. A parsing theory of word order universals. *Linguistic Inquiry* 21. 223–61. Online: https://www.jstor.org/stable/4178670.
HAWKINS, JOHN A. 1994. *A performance theory of order and constituency*. Cambridge: Cambridge University Press.
HAWKINS, JOHN A. 1998. Some issues in a performance theory of word order. *Constituent order in the languages of Europe*, ed. by Anna Siewierska, 729–81. Berlin: Mouton de Gruyter.
HAWKINS, JOHN A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.
HAWKINS, JOHN A. 2014. *Cross-linguistic variation and efficiency*. Oxford: Oxford University Press.
HAYS, DAVID G. 1964. Dependency theory: A formalism and some observations. *Language* 40.511–25. DOI: 10.2307/411934.
HERINGER, HANS JÜRGEN; BRUNO STRECKER; and RAINER WIMMER. 1980. *Syntax: Fragen-Lösungen-Alternativen*. Munich: Wilhelm Fink.
HOCHBERG, ROBERT A., and MATTHIAS F. STALLMANN. 2003. Optimal one-page tree embeddings in linear time. *Information Processing Letters* 87.59–66. DOI: 10.1016/S0020-0190(03)00261-8.
HOCKETT, CHARLES F. 1960. The origin of speech. *Scientific American* 203.88–96. Online: https://www.jstor.org/stable/24940617.
HOFMEISTER, PHILIP; PETER CULICOVER; and SUSANNE WINKLER. 2015. Effects of processing on the acceptability of 'frozen' extraposed constituents. *Syntax* 18.464–83. DOI: 10.1111/synt.12036.
HOFMEISTER, PHILIP; T. FLORIAN JAEGER; INBAL ARNON; IVAN A. SAG; and NEAL SNIDER. 2013. The source ambiguity problem: Distinguishing the effects of grammar and processing on acceptability judgments. *Language and Cognitive Processes* 28.48–87. DOI: 10.1080/01690965.2011.572401.
HOFMEISTER, PHILIP; LAURA STAUM CASASANTO; and IVAN A. SAG. 2014. Processing effects in linguistic judgment data: (Super-) additivity and reading span scores. *Language and Cognition* 6.111–45. DOI: 10.1017/langcog.2013.7.
HSU, ANNE S., and NICK CHATER. 2010. The logical problem of language acquisition: A probabilistic perspective. *Cognitive Science* 34.972–1016. DOI: 10.1111/j.1551-6709.2010.01117.x.
HUDSON, RICHARD A. 1984. *Word grammar*. Oxford: Blackwell.
HUDSON, RICHARD A. 1990. *English word grammar*. Oxford: Blackwell.
HUDSON, RICHARD A. 1995. Measuring syntactic difficulty. London: University College London, MS. Online: http://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf.
HUSAIN, SAMAR; SHRAVAN VASISHTH; and NARAYANAN SRINIVASAN. 2014. Strong expectations cancel locality effects: Evidence from Hindi. *PLOS ONE* 9:e100986. DOI: 10.1371/journal.pone.0100986.
IRURTZUN, ARITZ. 2009. Per què Y? Sobre la centralitat de la sintaxi a l'arquitectura de la gramàtica [Why Y? On the centrality of syntax in the architecture of grammar]. *Catalan Journal of Linguistics* 8.141–60. DOI: 10.5565/rev/catjl.145.
JAEGER, T. FLORIAN. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61.23–62. DOI: 10.1016/j.cogpsych.2010.02.002.
JAEGER, T. FLORIAN, and HARRY J. TILY. 2011. On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science* 2.323–35. DOI: 10.1002/wcs.126.
JOSHI, ARAVIND K.; KRISHNAMURTI VIJAY-SHANKER; and DAVID J. WEIR. 1991. The convergence of mildly context-sensitive grammar formalisms. *Foundational issues in natural language processing*, ed. by Peter Sells, Stuart Shieber, and Thomas Wasow, 31–81. Cambridge, MA: MIT Press.

Just, M. A., and P. A. Carpenter. 1992. A capacity theory of comprehension: Individual differences in working memory. *Psychological Review* 99.122–49. DOI: 10.1037/0033-295X.99.1.122.

Kayne, Richard S. 1994. *The antisymmetry of syntax*. Cambridge, MA: MIT Press.

Kiparsky, Paul. 2008. Universals constrain change; change results in typological generalizations. *Linguistic universals and language change*, ed. by Jeff Good, 23–53. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199298495.003.0002.

Kobele, Gregory Michael. 2006. *Generating copies: An investigation into structural identity in language and grammar*. Los Angeles: University of California, Los Angeles dissertation.

Konieczny, Lars. 2000. Locality and parsing complexity. *Journal of Psycholinguistic Research* 29.627–45. DOI: 10.1023/A:1026528912821.

Kuhlmann, Marco. 2013. Mildly non-projective dependency grammar. *Computational Linguistics* 39.355–87. DOI: 10.1162/COLI_a_00125.

Levy, Roger. 2005. *Probabilistic models of word order and syntactic discontinuity*. Stanford, CA: Stanford University dissertation.

Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106.1126–77. DOI: 10.1016/j.cognition.2007.05.006.

Levy, Roger. 2013. Memory and surprisal in human sentence comprehension. *Sentence processing*, ed. by Roger P. G. van Gompel, 78–114. Hove: Psychology Press.

Levy, Roger, and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. *Advances in Neural Information Processing Systems* 19.849–56. Online: https://papers.nips.cc/paper/3129-speakers-optimize-information-density-through-syntactic-reduction.

Liu, Haitao. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science* 9.159–91. Online: http://cogsci.snu.ac.kr/jcs/index.php/issues/?pageid=14&uid=76&mod=document.

Liu, Haitao. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua* 120.1567–78. DOI: 10.1016/j.lingua.2009.10.001.

Liu, Haitao; Chunshan Xu; and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews* 21.171–93. DOI: 10.1016/j.plrev.2017.03.002.

Marcus, Solomon. 1965. Sur la notion de projectivité. *Mathematical Logic Quarterly* 11.181–92. DOI: 10.1002/malq.19650110212.

McDonald, Janet L.; J. Kathryn Bock; and Michael H. Kelly. 1993. Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology* 25.188–230. DOI: 10.1006/cogp.1993.1005.

Mel'čuk, Igor A. 1988. *Dependency syntax: Theory and practice*. Albany: State University of New York Press.

Michaelis, Jens. 1998. Derivational minimalism is mildly context-sensitive. *Logical Aspects of Computational Linguistics (LACL 1998)*, 179–98. Berlin: Springer. DOI: 10.1007/3-540-45738-0_11.

Michaelis, Jens. 2001. Transforming linear context-free rewriting systems into minimalist grammars. *Logical Aspects of Computational Linguistics (LACL 2001)*, 228–44. Berlin: Springer. DOI: 10.1007/3-540-48199-0_14.

Michaelis, Jens, and Marcus Kracht. 1997. Semilinearity as a syntactic invariant. *Logical Aspects of Computational Linguistics (LACL 1996)*, 329–45. Berlin: Springer. DOI: 10.1007/BFb0052165.

Newmeyer, Frederick J. 1998. *Language form and language function*. Cambridge, MA: MIT Press.

Newmeyer, Frederick J. 2014. Where do motivations compete? *Competing motivations in grammar & usage*, ed. by Brian MacWhinney, Andrej Malchukov, and Edith Moravscik, 299–314. Oxford: Oxford University Press.

Nivre, Joakim. 2015. Towards a universal grammar for natural language processing. *Computational Linguistics and Intelligent Text Processing (CICLing 2015)*, 3–16. Cham: Springer. DOI: 10.1007/978-3-319-18111-0_1.

Nivre, Joakim; Željko Agić; Lars Ahrenberg; Lene Antonsen; Maria Jesus Aranzabe; Masayuki Asahara; Luma Ateyah; Mohammed Attia; Aitziber Atutxa;

LIESBETH AUGUSTINUS; et al. 2017. Universal dependencies 2.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics. Prague: Charles University. Online: http://hdl.handle.net/11234/1-2515.

NIVRE, JOAKIM, and JENS NILSSON. 2005. Pseudo-projective dependency parsing. *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics*, 99–106. DOI: 10.3115/1219840.1219853.

OSBORNE, TIMOTHY, and KIM GERDES. 2019. The status of function words in dependency grammar: A critique of Universal Dependencies (UD). *Glossa: a journal of general linguistics* 4(1):17. DOI: 10.5334/gjgl.537.

PARK, Y. ALBERT, and ROGER LEVY. 2009. Minimal-length linearizations for mildly context-sensitive dependency trees. *Proceedings of Human Language Technologies: The 2009 annual conference of the North American Chapter of the Association for Computational Linguistics*, 335–43. Online: http://www.aclweb.org/anthology/N/N09/N09-1038.

PIANTADOSI, STEVEN T., and EVELINA FEDORENKO. 2017. Infinitely productive language can arise from chance under communicative pressure. *Journal of Language Evolution* 2.141–47. DOI: 10.1093/jole/lzw013.

PITLER, EMILY; SAMPATH KANNAN; and MITCHELL MARCUS. 2013. Finding optimal 1-endpoint-crossing trees. *Transactions of the Association for Computational Linguistics* 1. 13–24. DOI: 10.1162/tacl_a_00206.

POLLARD, CARL, and IVAN A. SAG. 1987. *Information-based syntax and semantics*. Stanford, CA: CSLI Publications.

POPEL, MARTIN; DAVID MAREČEK; JAN ŠTĚPÁNEK; DANIEL ZEMAN; and ZDENĚK ŽABOKRTSKÝ. 2013. Coordination structures in dependency treebanks. *Proceedings of the 51st annual meeting of the Association for Computational Linguistics*, 517–27. Online: https://www.aclweb.org/anthology/P13-1051.

POPP, EARL Y., and MICHAEL J. SERRA. 2016. Adaptive memory: Animacy enhances free recall but impairs cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 42.186–201. DOI: 10.1037/xlm0000174.

PRINCE, ELLEN F. 1981. Toward a taxonomy of given/new information. *Radical pragmatics*, ed. by Peter Cole, 223–55. New York: Academic Press.

RAJKUMAR, RAJAKRISHNAN; MARTEN VAN SCHIJNDEL; MICHAEL WHITE; and WILLIAM SCHULER. 2016. Investigating locality effects and surprisal in written English syntactic choice phenomena. *Cognition* 155.204–32. DOI: 10.1016/j.cognition.2016.06.008.

RIJKHOFF, JAN. 1986. Word order universals revisited: The principle of head proximity. *Belgian Journal of Linguistics* 1.95–125. DOI: 10.1075/bjl.1.05rij.

RIJKHOFF, JAN. 1990. Explaining word order in the noun phrase. *Linguistics* 28.5–42. DOI: 10.1515/ling.1990.28.1.5.

ROS, IDOIA; MIKEL SANTESTEBAN; KUMIKO FUKUMORA; and ITZIAR LAKA. 2015. Aiming at shorter dependencies: The role of agreement morphology. *Language, Cognition and Neuroscience* 30.1156–74. DOI: 10.1080/23273798.2014.994009.

SCONTRAS, GREGORY; JUDITH DEGEN; and NOAH D. GOODMAN. 2019. On the grammatical source of adjective ordering preferences. *Semantics and Pragmatics* 12:7. DOI: 10.3765/sp.12.7.

SHALIZI, COSMA ROHILLA, and JAMES P. CRUTCHFIELD. 2001. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics* 104.817–79. DOI: 10.1023/A:1010388907793.

SHANNON, CLAUDE E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27.623–56. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

SHIEBER, STUART M. 1985. Evidence against the context-freeness of natural language. *The formal complexity of natural language*, ed. by Walter J. Savitch, Emmon Bach, William Marsh, and Gila Safran-Naveh, 320–34. Dordrecht: Springer. DOI: 10.1007/978-94-009-3401-6_12.

SHIH, STEPHANIE, and JASON GRAFMILLER. 2011. Weighing in on end weight. Paper presented at the 85th annual meeting of the Linguistic Society of America, Pittsburgh, PA.

SHIH, STEPHANIE; JASON GRAFMILLER; RICHARD FUTRELL; and JOAN BRESNAN. 2015. Rhythm's role in the genitive construction choice in spoken English. *Rhythm in phonetics, grammar, and cognition*, ed. by Ralf Vogel and Reuben van de Vijver, 208–34. Berlin: De Gruyter Mouton.

Silverstein, Michael. 1976. Hierarchy of features and ergativity. *Grammatical categories in Australian languages*, ed. by R. M. W. Dixon, 112–71. Canberra: Australian Institute of Aboriginal Studies.
Sleator, Daniel, and David Temperley. 1991. Parsing English with a link grammar. Computer science technical report CMU-CS-91-196. Pittsburgh, PA: Carnegie Mellon University. Online: https://www.link.cs.cmu.edu/link/ftp-site/link-grammar/LG-tech-report.pdf.
Slobin, Dan I. 1973. Cognitive prerequisites for the development of grammar. *Studies of child language development*, ed. by Dan I. Slobin and Charles A. Ferguson, 175–209. New York: Holt, Rinehart & Winston.
Smith, Nathaniel J., and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition* 128.302–19. DOI: 10.1016/j.cognition.2013.02.013.
Stabler, Edward P. 1997. Derivational minimalism. *Logical Aspects of Computational Linguistics (LACL 1996)*, 68–95. Berlin: Springer. DOI: 10.1007/BFb0052152.
Temperley, David. 2005. The dependency structure of coordinate phrases: A corpus approach. *Journal of Psycholinguistic Research* 34.577–601. DOI: 10.1007/s10936-005-9165-2.
Temperley, David. 2007. Minimization of dependency length in written English. *Cognition* 105.300–333. DOI: 10.1016/j.cognition.2006.09.011.
Temperley, David, and Daniel Gildea. 2018. Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics* 4.67–80. DOI: 10.1146/annurev-linguistics-011817-045617.
Tesnière, Lucien. 1959. *Eléments de syntaxe structurale*. Paris: Librairie C. Klincksieck.
Vasishth, Shravan; Nicolas Chopin; Robin Ryder; and Bruno Nicenboim. 2017. Modelling dependency completion in sentence comprehension as a Bayesian hierarchical mixture process: A case study involving Chinese relative clauses. arXiv:1702.00564 [stat.AP]. Online: https://arxiv.org/abs/1702.00564.
Vennemann, Theo. 1974. Theoretical word order studies: Results and problems. *Papiere zur Linguistik* 7.5–25.
von der Gabelentz, Georg. 1901. *Die Sprachwissenschaft: Ihre Aufgaben, Methoden, und bisherigen Ergebnisse*. Leipzig: Weigel.
Wasow, Thomas. 2002. *Postverbal behavior*. Stanford, CA: CSLI Publications.
Weir, David Jeremy. 1988. *Characterizing mildly context-sensitive grammar formalisms*. Philadelphia: University of Pennsylvania dissertation.
Xu, Chunshan, and Haitao Liu. 2015. Can familiarity lessen the effect of locality? A case study of Mandarin Chinese subjects and the following adverbials. *Poznań Studies in Contemporary Linguistics* 51.463–85. DOI: 10.1515/psicl-2015-0018.
Yadav, Himanshu; Samar Husain; and Richard Futrell. 2019. Are formal restrictions on crossing dependencies epiphenomenal? *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theory*, 2–12. DOI: 10.18653/v1/W19-7802.
Yamashita, Hiroko, and Franklin Chang. 2001. 'Long before short' preference in the production of a head-final language. *Cognition* 81.B45–B55. DOI: 10.1016/S0010-0277(01)00121-4.
Yngve, Victor H. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society* 104.444–66. Online: https://www.jstor.org/stable/985230.
Yu, Xiang; Agnieszka Falenska; and Jonas Kuhn. 2019. Dependency length minimization vs. word order constraints: An empirical study on 55 treebanks. *Proceedings of the First International Conference on Quantitative Syntax*. Online: https://www.aclweb.org/anthology/W19-7911.pdf.
Zipf, George Kingsley. 1936. *The psychobiology of language*. London: Routledge.
Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort*. Oxford: Addison-Wesley.

[rfutrell@uci.edu]
[rplevy@mit.edu]
[egibson@mit.edu]

[Received 22 March 2019;
revision invited 14 July 2019;
revision received 30 September 2019;
revision invited 20 November 2019;
revision received 17 December 2019;
accepted 28 December 2019]