COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 45 (2021) e13056

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13056

Word Order Variation is Partially Constrained by Syntactic Complexity

Yingqi Jing,[a,b,c] Paul Widmer,[a,b] Balthasar Bickel[a,b]

[a] *Department of Comparative Language Science, University of Zurich*
[b] *Center for the Interdisciplinary Study of Language Evolution, University of Zurich*
[c] *Department of Linguistics and Philology, Uppsala University*

Received 21 December 2020; received in revised form 21 July 2021; accepted 14 September 2021

Abstract

Previous work suggests that when speakers linearize syntactic structures, they place longer and more complex dependents further away from the head word to which they belong than shorter and simpler dependents, and that they do so with increasing rigidity the longer expressions get, for example, longer objects tend to be placed further away from their verb, and with less variation. Current theories of sentence processing furthermore make competing predictions on whether longer expressions are preferentially placed as early or as late as possible. Here we test these predictions using hierarchical distributional regression models that allow estimates of word order and word order variation at the level of individual dependencies in corpora from 71 languages, while controlling for confounding effects from the type of dependency (e.g., subject vs. object), and the type of clause (main vs. subordinate) involved as well as from trends that are characteristic of individual languages, language families, and language contact areas. Our results show the expected correlations of length with position and variation only for two out of six dependency types (obliques and nominal modifiers) and no difference between clause types. These findings challenge received theories of across-the-board effects of complexity on word order and word order variation and call for theoretical models that relativize effects to specific kinds of syntactic structures and dependencies.

Keywords: Word order variation; Dependency length; Heaviness serialization principle; Distributional regression model; Dependency treebanks

[Correction added on 07 May, 2022 after first online publication: CSAL funding statement has been added.]

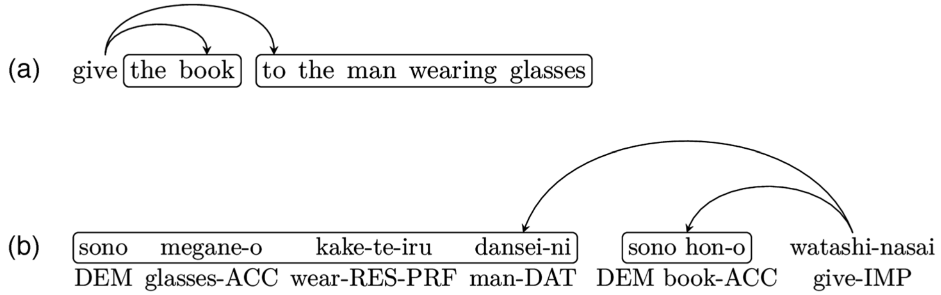
Correspondence should be sent to Yingqi Jing, Department of Linguistics and Philology, Uppsala University, 75126 Uppsala, Sweden. E-mail:yingqi.jing@lingfil.uu.se

1. Introduction

Our communication channels (speech, gesture, writing) impose a strict linearization of syntactic structure in language. While linearizations differ widely across languages (e.g., English puts the verb before its object, Japanese puts it after the object) (Greenberg, 1963), there is a growing body of evidence that they are partially shaped by general principles of efficiency and economy that facilitate the processing, planning, and learning of language (Culbertson, Smolensky, & Legendre, 2012; Dryer, 1992; Fedzechkina, Chu, & Florian Jaeger, 2018; Futrell, Levy, & Gibson, 2020; Gennari, Sloman, Malt, & Fitch, 2002; Gibson, 1998; Gibson et al., 2019; Hall, Mayberry, & Ferreira, 2013; Hawkins, 1983, 2014; MacDonald, 2013). While current proposals differ in their more specific predictions, they agree in the general hypothesis that the more complex a syntactic element is, the more rigidly it is placed at the periphery of a dependency structure, that is it is more rigidly final or more rigidly initial. While syntactic complexity can be measured in various ways, a useful approximation has proven to be the sheer length of a dependent or constituent in the number of word-like segments.

To illustrate the intuition behind this idea, consider the following examples from English and Japanese, where directed arcs denote which element depends on which other element in terms of "theme-of" and "recipient-of" dependencies. While the two languages differ in where they place the verb (initial in English, final in Japanese), both prefer to place the longer and more complex dependent ("to the man wearing glasses") at the periphery of the verb's dependency structure. But when both arguments are shorter and simpler, they are more freely expressed as either "give the man the book" or "give the book to the man."

(1) English and Japanese ditransitive examples


(a)
give the book to the man wearing glasses
(b)
sono megane-o kake-te-iru dansei-ni sono hon-o watashi-nasai
DEM glasses-ACC wear-RES-PRF man-DAT DEM book-ACC give-IMP

The hypothesis that longer dependents are preferentially placed in the periphery is primarily motivated by more general observations of dependency length minimization (DLM) (Ferrer i Cancho, 2004; Futrell et al., 2020; Gibson, 1998, 2000; Jing, Blasi, & Bickel, in press; Liu, 2008; Temperley, 2008). The core finding of DLM research is that speakers seek to minimize the total length of dependencies in a given structure because this reduces pressure on working memory and keeps adjacent information that belongs together (Choi, 2007; Faghiri & Samvelian, 2014, 2020; Futrell et al., 2020; Gibson et al., 2019; Hawkins, 1994; Ros, Santesteban, Fukumura, & Laka, 2015; Yamashita, 2002; Yamashita & Chang, 2001). One specific effect of this bears on the peripheral placement of long dependents:

when dependents occur on the same side of their head, placing the shortest dependent closest to the verb reduces the total length of dependencies. Thus, in both English and Japanese, placing the shorter theme ("the book") closer to the verb ("give") than the recipient ("the man wearing glasses") ensures that the total distance between the verb and its dependents is the shortest possible. If the recipient was placed closer to the verb ("give the man wearing glasses the book"), this would lengthen the dependency between the verb and the theme ("the book") in both languages. These specific effects of DLM are best ensured when linearization is not entirely left to speakers' contextual needs, but when grammatical rules or constraints evolve in such a way that they make the peripheral placement of longer dependents relatively rigid and allow only relatively short dependents to vary in their positions. Such rules or constraints will generate more dependencies with shorter overall distances than dependencies with longer overall distances (Gildea & Temperley, 2010; Temperley, 2008).

The associations of long dependents with rigid placement and of short dependents with variable placement have often been noted in the linguistic literature, for example, in terms of a "mobility" principle that predicts short dependents to be more flexible than long dependents (Hawkins, 1983). One effect of the mobility principle is that it can disrupt the oft-noted harmony in head/dependent ordering across different types of dependencies (Culbertson et al., 2012; Dryer, 1992; Greenberg, 1963; Lehmann, 1973; Temperley, 2008; Vennemann, 1974). For example, the order of verbs and their object ("see the house") correlates with the order of nouns and dependent nominal modifiers ("the owner of the house"). The order of nouns and adjectives, however, appears to escape the correlation when adjectives are short, single-word expressions, but not when they are fully fledged adjective phrases (Dryer, 1992). For example, the English noun phrases with a single adjective do not follow the general head-initial (HI) preference of the language (cf. head-final [HF] "rich countries") while noun phrases with fully fledged adjective phrases do (cf. HI "countries rich in resources"). Gulordava (2018) shows that this is a general pattern in corpus data across Romance languages.

While these considerations predict a peripheral and rigid placement of more complex dependents, they do not make a prediction on whether the periphery is the first or the last position in a dependency structure. On this question, there are two competing theories. One theory goes back to Behaghel's "law of growing parts" and privileges the last position, following a general short/simple before long/complex linearization (Behaghel, 1909). Behaghel's law has been recast in various ways in the literature, such as in terms of an end-weight preference (Quirk, 1972; Wasow, 1997), the heaviness serialization principle (Hawkins, 1983), a heavy NP shift (Arnold, Losongco, Wasow, & Ginstrom, 2000), a short-before-long ordering (Fedzechkina et al., 2018; Temperley, 2008), and an easy first bias (MacDonald, 2013). Behaghel's law is commonly motivated by the cognitive simplicity of short elements (Bock, 1982; Bock & Levelt, 1994; Christianson & Ferreira, 2005; Levelt, 1999; MacDonald, 2013). This simplicity effect is potentially strengthened by the fact that short elements (such as pronouns) tend to represent easy-to-access, known information (Arnold et al., 2000; J. M. Gundel, 1975; J. K. Gundel, 1988; Haviland & Clark, 1974; Krifka, 2008; MacDonald, 2013). For pairs of heads and dependents (where the head is usually a single word), Behaghel's law predicts that longer dependents tend to be placed after their head (as they are in the English example in 1a), while short dependents are expected to be more variably placed on either side of the head (as is the case of some English adverbs which can be ordered both as in "she quickly finished" or as in "she finished quickly") (Gildea & Temperley, 2010; Gildea & Jaeger, 2015; Temperley, 2008). In other words, across languages, one would expect more HI and less variable structures with longer and more complex dependents (long→head-initial→less variable).

The other theory privileges the first position and is based on the notion that speakers seek to convey new and rich information as early as possible. This has been motivated by a general decrease in communicative importance over the course of an utterance since short elements (e.g., pronouns) represent information that is less important and can be inferred from the previous discourse (Givón, 1988; J. K. Gundel, 1988). An alternative motivation is grounded in expectation-based theories, which propose that the early placement of semantically rich information reduces surprisal and facilitates the processing of the upcoming head by increasing its predictability (Husain, Vasishth, & Srinivasan, 2014; Konieczny, 2000; R. Levy, 2008; R. P. Levy & Keller, 2013; Vasishth & Lewis, 2006), and they also facilitate sentence planning by keeping plans distinct from alternatives (Sauppe et al., 2021). For example, a complex object noun phrase facilitates the processing of a final verb in cases like the following:

(2) Expectation-based facilitation in German (Konieczny 2000)
a. Er hat das Buch, [das Lisa gestern gekauft hatte], hingelegt.
he has the book that Lisa yesterday bought had laid.down
b. Er hat das Buch hingelegt, [das Lisa gestern gekauft hatte].
he has the book laid.down that Lisa yesterday bought had
'He has laid down the book that Lisa had bought yesterday.'

In self-paced reading, the verb is read faster in such examples when it follows the semantically rich and complex noun phrase in 2a ("the book that Lisa bought yesterday") than when it follows the short object noun phrase "the book" in 2b, where the relative clause is extraposed (Konieczny, 2000). Since these facilitation effects matter most for long and information-rich dependents, one would expect less variation in placement with longer dependencies (long→head-final→less variable).

In what follows, we use cross-linguistic corpus data to assess the evidence for the overall hypothesis that longer dependents tend to be placed at the periphery of a dependency structure and to evaluate the relative evidence for the more specific hypotheses of a cross-linguistic preference for early versus late placements of longer dependents. We do this by statistically modeling the order and variation of where dependents are placed in relation to their length in individual dependency structures, controlling for general trends in languages, families, and geographical areas. We furthermore control for differences arising from types of clauses and types of dependencies. With regard to clause types, it has been hypothesized that word order varies more in main than in dependent clauses (Ross, 1973), and that, consistently with this, main clauses are more amenable to variation over time (Givón, 1979; Hock, 1991; Lightfoot, 1982; Vennemann, 1975). This hypothesis is generally motivated by the assumption that subordinate clauses tend to contain presupposed background information (Givón, 1979; Hooper & Thompson, 1973) or that they tend to be processed as whole chunks (Bybee, 2002). With regard to dependency types, it has been noted that they differ in the variation they tend to tolerate. For instance, overall, word orders in noun modifiers tend to have less flexible order than core verbal arguments (Gulordava & Merlo, 2015; Tily, 2010), and oblique arguments are found to be the most variable (Levshina, 2019).

Our focus is entirely on individual dependencies, and we seek to assess the proposed general principles that push longer dependents to early or late in a given dependency and that they show less variation than short dependents. Therefore, we do not presently explore potential interactions of a given dependency with other dependencies in the same sentence or the same language, nor do we address effects from the combination of several dependencies into complex structures. As a consequence, our study differs from DLM research, where the total lengths over all dependencies are compared to randomized baselines generated by entire grammars. A randomization-based approach is not readily suitable for simultaneous measurements of order and variation at the level of individual dependencies instead of entire grammars.

2. Data and methods

2.1. Data

The corpus data in this study comes from the Universal Dependencies (UD) database, version 2.7 (Nivre et al., 2020). The UD database provides a collection of dependency-annotated corpora of diverse languages in the general framework of Dependency Grammar (de Marneffe & Manning, 2008; de Marneffe & Nivre, 2019; Hudson, 1984; Liu, 2009; Melčuk, 1988; Petrov, Das, & McDonald, 2012; Zeman, 2008). Each pair of words in a sentence is linked via directed arcs that indicate their head directions and dependency types. Given continued controversies about whether functional categories, such as determiners or auxiliary verbs, should be considered as heads or dependents(Futrell et al., 2020; Groß & Osborne, 2015; Osborne & Gerdes, 2019; Osborne & Maxwell, 2015), we focus exclusively on dependencies between major lexical categories, such as V(erb), N(oun), Adj(ective), and Adv(erb). These have a clear head/dependent structure and are cross-linguistically more comparable than functional categories. Furthermore, the position of pronouns and auxiliaries is often subject to confounding factors from phonology (as when auxiliaries need a host for cliticization).

This gives us a set of six dependencies (verb/nominal subject (*the boy ran away*), verb/nominal object (*read the book*), verb/oblique (*read the book [over the weekend]*), verb/adverbial modifier (*read silently*), noun/nominal modifier (*my friend's book*), and noun/adjectival modifier (*a new book*)) in 71 treebanks from UD v2.7. The choice of these dependencies guarantees a sufficient number of tokens in each corpus and across main and subordinate clauses ($n \geq 100$), and a minimum coverage of 40 languages. For those languages with more than one treebank, we choose the largest one. The information of these treebanks is summarized in Table 1.

Table 1
Overview of 71 dependency treebanks from UD v2.7. We simplify the geographical distribution as Europe versus not since most languages are from Europe (see Fig. S4 for the geographic distribution of languages)

Language	Family	Europe	Sentences	Word Token	Language	Family	Europe	Sentences	Word token
Afrikaans	Indo-European	F	1,934	49,276	Irish	Indo-European	T	4,910	115,969
Akkadian	Afro-Asiatic	F	1,804	21,962	Italian	Indo-European	T	14,167	298,343
Ancient Greek	Indo-European	T	17,080	213,999	Japanese	Japanese	F	57,028	1,250,875
Arabic	Afro-Asiatic	F	19,738	738,889	Korean	Korean	F	27,363	350,090
Armenian	Indo-European	T	2,502	52,630	Latin	Indo-European	T	26,977	450,515
Bambara	Mande	F	1,026	13,823	Latvian	Indo-European	T	13,643	219,955
Basque	Basque	T	8,993	121,443	Lithuanian	Indo-European	T	3,642	70,047
Belarusian	Indo-European	T	23,534	275,153	Maltese	Afro-Asiatic	F	2,074	44,162
Breton	Indo-European	T	888	10,054	Naija	Creole	F	9,242	140,729
Bulgarian	Indo-European	T	11,138	156,149	North Sami	Uralic	T	3,122	26,845
Buryat	Mongolic	F	927	10,185	Norwegian	Indo-European	T	20,044	310,221
Cantonese	Sino-Tibetan	F	1,004	13,918	Old Church Slavonic	Indo-European	T	6,338	57,563
Catalan	Indo-European	T	16,678	531,971	Old French	Indo-European	T	17,678	170,740
Chinese	Sino-Tibetan	F	4,997	123,291	Old Russian	Indo-European	T	16,944	149,780

(*Continued*)

Table 1
Continued

Language	Family	Europe	Sentences	Word Token	Language	Family	Europe	Sentences	Word token
Classical Chinese	Sino-Tibetan	F	48,434	233,122	Persian	Indo-European	F	29,107	501,776
Coptic	Afro-Asiatic	F	1,873	48,631	Polish	Indo-European	T	22,152	350,036
Croatian	Indo-European	T	9,010	199,409	Portuguese	Indo-European	T	12,078	319,853
Czech	Indo-European	T	87,913	1,506,484	Romanian	Indo-European	T	26,225	572,436
Danish	Indo-European	T	5,512	100,733	Russian	Indo-European	T	61,889	1,106,296
Dutch	Indo-European	T	13,578	208,470	Sanskrit	Indo-European	F	3,997	27,117
English	Indo-European	T	16,622	254,829	Scottish Gaelic	Indo-European	T	3,173	60,417
Erzya	Uralic	T	1,690	17,148	Serbian	Indo-European	T	4,384	97,673
Estonian	Uralic	T	30,972	437,769	Slovak	Indo-European	T	10,604	106,043
Faroese	Indo-European	T	1,621	40,484	Slovenian	Indo-European	T	8,000	140,670
Finnish	Uralic	T	15,136	202,194	Spanish	Indo-European	T	17,680	549,569
French	Indo-European	T	18,535	573,370	Swedish	Indo-European	T	6,026	96,819
Galician	Indo-European	T	3,993	138,837	Tamil	Dravidian	F	600	9,581
German	Indo-European	T	189,928	3,399,390	Turkish German	Mixed	F	1,891	31,946
Gothic	Indo-European	T	5,401	55,336	Turkish	Turkic	F	9,761	122,383
Greek	Indo-European	T	2,521	63,441	Ukrainian	Indo-European	T	7,060	122,091
Hebrew	Afro-Asiatic	F	6,216	161,417	Upper Sorbian	Indo-European	T	646	11,196
Hindi English	Mixed	F	1,898	26,909	Urdu	Indo-European	F	5,130	138,077
Hindi	Indo-European	F	16,647	351,704	Vietnamese	Austro-Asiatic	F	3,000	43,754
Hungarian	Uralic	T	1,800	42,032	Welsh	Indo-European	T	1,657	32,911
Icelandic	Indo-European	T	44,029	985,057	Wolof	Niger-Congo	F	2,107	44,258
Indonesian	Austronesian	F	5,593	121,923					

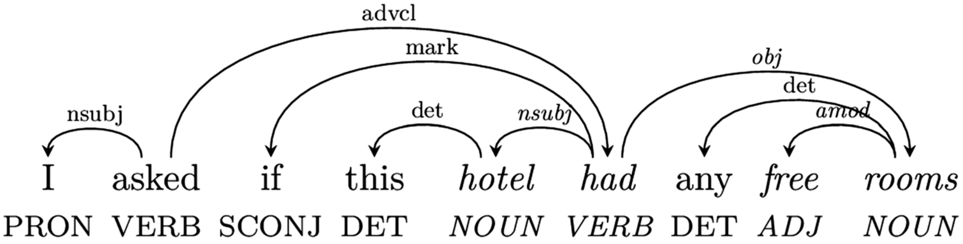

advcl
mark
obj
det
nsubj
det
nsubj
amod
I asked if this hotel had any free rooms
PRON VERB SCONJ DET NOUN VERB DET ADJ NOUN

Fig. 1. Dependency graph of an English sentence with an adverbial clause: each terminal is annotated by its parts of speech tag (e.g., PRON for "pronoun"), and each arc for its dependency types (e.g., "nsubj" for "nominal subject"). The directed arcs go from the head to its dependents, indicating the head directions, for example, the subject "hotel" precedes the verbal head "had" forming a HF dependency, and the object "rooms" forms a HI dependency with the governor "had."

The dependencies in UD are defined by combining semantic and syntactic criteria[1]: a nominal subject ("nsubj") is the most agent-like argument in an active transitive or the single argument in an intransitive clause, and an object of the verb ("obj") is the most patient-like argument in a transitive clause. The oblique relation ("obl") subsumes both obliquely marked arguments (e.g., adpositional objects like "to the reader" in *give it to the reader* or "by"-agents in passives) and adjuncts (e.g., time or location expressions like "in the city"). An adverbial modifier ("advmod") is an adverb or adverbial phrase that serves to modify a predicate, and a nominal modifier ("nmod") is used for nominal dependents of another noun or noun phrase and functionally corresponds to an attribute or genitive complement. An adjectival modifier ("amod") is an adjectival word or phrase that modifies a noun.

For illustration, Fig. 1 provides an annotated English sentence containing a dependent clause, where the target dependencies are displayed in italic. We strictly separate the word orders in main clauses, including only simple clauses, and in subordinate clauses, including adjectival (relative) clauses ("acl"), adverbial clauses ("advcl"), clausal subjects ("csubj"), and clausal complements ("ccomp"). Accordingly, the main clause part in Fig. 1 is also excluded to avoid any potential confounds from the overall the complexity of the sentence.

We measure dependency length (DL) as the linear distance between each pair of a head and its dependents in numbers of words (as defined by the UD word segmentation). For example, the adjacent dependency between the verb *had* and nominal subject *hotel* yields a length of 1, and the DL between the head verb *had* and object *rooms* is 3. The adjective modifier *free* is adjacent to its head noun *rooms* with a DL of 1.

2.2. Methods

Previous corpus-based studies on word order variation mostly use entropy to quantify word order freedom at the language level. For example, Futrell, Mahowald, and Gibson (2015) use entropy to quantify free ordering of subjects and objects as a predictor of case marking. Gulordava (2018) takes entropy as a measure of a diachronic change towards more rigid word order in the history of Greek. Levshina (2019) uses entropy to show that function elements generally show the least degree of word order variation, followed by noun modifiers, core verbal arguments, and adjuncts/obliques.

However, since we seek to estimate effects of length within single dependency structures, entropies would have to be estimated for each length, or each bin of lengths, over the corpus of a given language. Since lengths show a power-law distribution, this severely reduces sample sizes for longer lengths, making estimates unstable, reducing statistical power, and introducing problematic associations between sample size and entropy (Chao, Wang, & Jost, 2013; DeDeo, Hawkins, Klingenstein, & Hitchcock, 2013). Moreover, measuring entropies at the corpus (language) level risks underestimating the actual variation within it, since entropies based on observed frequencies may have a poor representation of the real distributions, especially for rare dependency types or in small corpora (see Futrell et al., 2015, for bootstrap estimators of entropy using subcorpora and complete corpora). It is in fact challenging to get a reliable estimate of entropy in high-dimensional and undersampled datasets, and some bias correction is often needed (Chao et al., 2013; DeDeo et al., 2013; Hausser & Strimmer, 2009; Nemenman, Shafee, & Bialek, 2002; Wolpert & Wolf, 1995).

In response to this, we move away from entropy measures of word order flexibility. As an alternative, we adopt distributional (location/scale) regression models to simultaneously predict head directionality (location) and variation (scale) of different word orders. The distributional models also allow us to directly evaluate the role of each predictor (DL, dependency type, clause type) on both word order preferences and their variation at both population ("fixed effects") and language ("random effects") levels (Nalborczyk, Batailler, Lœvenbruck, Vilain, & Bürkner, 2019). The distributional models relax the assumption of a constant scale parameter, but this comes with a degree of complexity that makes these models hard to fit in a frequentist maximum-likelihood approach. In response to this, we fit the models in a Bayesian framework. The development of Markov chain Monte Carlo (MCMC) algorithms, especially the Hamiltonian Monte Carlo and its variant No-U-Turn Sampler, makes it possible to efficiently sample from high-dimensional distributions and estimate parameters in a complex distributional model.

We predict the direction and variation of each word order in a model assuming a beta-binomial likelihood function. The beta-binomial distribution is a mixture of a binomial and a beta distribution. It generalizes the binomial distribution, and can capture overdispersion. With the beta-binomial model, the binomial probability is randomly drawn from a beta distribution $\mathbf{B}(\alpha, \beta)$ with hyperparameters $\alpha > 0$ and $\beta > 0$.

$$\text{Beta2}(\mu, \phi) = \text{Beta}\left(\alpha = \frac{\mu}{\phi}, \beta = \frac{(1-\mu)}{\phi}\right).$$

To make it easier to interpret the results, we choose a slightly different parameterization by using the parameters of a mean ($\mu = \frac{\alpha}{\alpha+\beta}$) for the location and the reciprocal of the precision parameter ($\phi = \frac{1}{\alpha+\beta}$) for the scale [cf. (Bürkner, 2018; Jørgensen, 1997; Kruschke, 2014) for similar parameterizations]. The parameter μ represents the expected binomial probability of each word order, and ϕ is the dispersion parameter indicating the degree of variation of word orders. This model allows us to predict the probability of each word order (H[ead]I[nitial] or H[ead]F[inal]) and its variation at the same time and given the same covariates.

Table 2
Models for estimating head-finality of each word order (μ) and word order variation (ϕ) using `brms`-style notation. CLS: clause type; DEP: dependency type; DL: dependency length; lang: language; fam: family. The family and area groups control for phylogenetic and spatial autocorrelation, respectively

Name	Regression
Intercept	$\mu \sim 1 + (1 \mid \text{lang}) + (1 \mid \text{fam}) + (1 \mid \text{area})$
	$\phi \sim 1 + (1 \mid \text{lang}) + (1 \mid \text{fam}) + (1 \mid \text{area})$
Clause type	$\mu \sim 1 + \text{CLS} + (1 + \text{CLS} \mid \text{lang}) + (1 \mid \text{fam}) + (1 \mid \text{area})$
	$\phi \sim 1 + \text{CLS} + (1 + \text{CLS} \mid \text{lang}) + (1 \mid \text{fam}) + (1 \mid \text{area})$
Dependency type	$\mu \sim 1 + \text{DEP} + (1 + \text{DEP} \mid \text{lang}) + (1 \mid \text{fam}) + (1 \mid \text{area})$
	$\phi \sim 1 + \text{DEP} + (1 + \text{DEP} \mid \text{lang}) + (1 \mid \text{fam}) + (1 \mid \text{area})$
DL	$\mu \sim 1 + \text{DL} + (1 + \text{DL} \mid \text{lang}) + (1 \mid \text{fam}) + (1 \mid \text{area})$
	$\phi \sim 1 + \text{DL} + (1 + \text{DL} \mid \text{lang}) + (1 \mid \text{fam}) + (1 \mid \text{area})$
Dependency type by clause type	$\mu \sim 1 + \text{DEP} + (1 + \text{DEP} \mid \text{CLS}) + (1 + \text{DEP} \mid \text{lang}) + (1 \mid \text{fam}) + (1 \mid \text{area})$
	$\phi \sim 1 + \text{DEP} + (1 + \text{DEP} \mid \text{CLS}) + (1 + \text{DEP} \mid \text{lang}) + (1 \mid \text{fam}) + (1 \mid \text{area})$
DL by clause type	$\mu \sim 1 + \text{DL} + (1 + \text{DL} \mid \text{CLS}) + (1 + \text{DL} \mid \text{lang}) + (1 \mid \text{fam}) + (1 \mid \text{area})$
	$\phi \sim 1 + \text{DL} + (1 + \text{DL} \mid \text{CLS}) + (1 + \text{DL} \mid \text{lang}) + (1 \mid \text{fam}) + (1 \mid \text{area})$
DL by dependency type	$\mu \sim 1 + \text{DL} + (1 + \text{DL} \mid \text{DEP}) + (1 + \text{DL} \mid \text{lang}) + (1 \mid \text{fam}) + (1 \mid \text{area})$
	$\phi \sim 1 + \text{DL} + (1 + \text{DL} \mid \text{DEP}) + (1 + \text{DL} \mid \text{lang}) + (1 \mid \text{fam}) + (1 \mid \text{area})$

We build and compare distributional regression models as shown in Table 2. At the population level ("fixed effects"), we estimate coefficients ofDL, clause type (CLS), and dependency type (DEP), allowing for group-level adjustments per language ("lang"), language family ("fam"), and area. This model structure can not only capture the language-specific variation but also allow us to compare the effect of DL on word orders between population level and language level.

We model interactions between DL and types of clauses and dependencies by allowing slopes to vary across types. We opt for this strategy (rather than models with population-level interactions) in order to profit from the shrinkage effects that arise when information is partially pooled across types (McElreath, 2020; Nalborczyk et al., 2019). This strategy avoids overfitting at the population level, and at the same time it accommodates the unequal and sometimes small sample sizes that arise when estimating effects across particular combinations of lengths and types.

In these models, the μ response is determined by the observed number of HF dependencies in the total number of n dependencies at a given combination of levels in the terms of the model. It is linked to the expected binomial probability of head-finality for each word order via the log odds (logit) function. The ϕ response is the dispersion parameter for word order variation. Higher ϕ values suggest more word order variation, while lower ϕ values suggest more rigid word order.

We treat DL as a monotonic predictor, *mo(DL)*, which accommodates non-linearity and is specifically designed for integer predictors and ordered factors more generally (Bürkner & Charpentier, 2020). The parameterization of monotonic effects is realized by joint estimations of the average word order differences due to DL and the relative proportions of observed differences between two adjacent categories. If the word order differences are evenly distributed between all adjacent levels of DL, a linear relationship of DL on word orders is guaranteed. Otherwise, a non-linear relationship is achieved. Thus, the *mo(DL)* term can not only evaluate the overall effect of DL on word orders but also identify the most influential region between consecutive levels of DL.

We fit the models in `Stan` (Stan Development Team, 2020) via the `brms` (Bürkner, 2017) and `cmdstanr` (Gabry & Češnovar, 2020) interfaces to `R` (R Core Team, 2020). We choose weakly informative priors by setting *Student-t(20, 0, 1)* for intercepts and slopes, and *half-Cauchy(0, 1)* for the standard deviations. For each model, we run two independent MCMC chains for 3,000 iterations with a warm-up of 1,000 iterations. We check for the convergence by ensuring effective sample sizes $ESS > 100$ and diagnostic $\hat{R} < 1.05$ for each parameter.

We assess the performance of models by leave-one-out cross-validation in terms of the expected log pointwise predictive density (*elpd*). We approximate the *eldp* through Pareto-smoothed importance sampling (PSIS) from the posterior log probabilities (McElreath, 2020; Yao et al., 2018). Apart from examining raw *eldp* estimates, we furthermore compare model performance by model stacking, that is by allocating weights to models in such a way that they jointly maximize prediction quality (Vehtari et al., 2020).

To get the posterior predictive effects for each predictor, we furthermore build a full model by including all predictors (DL, dependency types, and clause types) and their mutual interactions.[2.] We report 90% credible intervals since they are computationally more stable than wider intervals when the effective sample size is below 10k (Kruschke, 2014).

Code and data summaries are available in the Supporting Information; executable code and input data are accessible at https://osf.io/4s67a/.

3. Results

Fig. 2 visualizes the relationship between DL and word order across clause and dependency types in the raw data, using medians to capture the proportion of HF orders and boxplot quartiles to capture the variation in that proportion. The plot suggests that dependency type, but not clause type, has an appreciable impact on the relationship between DL and word order. These impressions are confirmed by the results from model comparison where dependency type turns out to be the most important predictor of word order, by itself (leveraging 46% of the total weight; Table 3) or as a source of variation for the effect of DL on word order (15%). Clause type is relevant for prediction only when modeled as a source of variation for the effect of dependency type (38%). Consistent with this, the posterior probability of population-level differences between clause types includes 0 even in its 60% credible interval (60% CI $= [-0.04, 0.3]$ for head direction; 60% CI $= [-0.35, 0.001]$ for word order variation). In what follows, we focus on the conditional effects of DL across dependency types in the full model, which subsumes all models from Table 3.

We first present the direction estimates (Fig. 3) and then the variation estimates (Fig. 4). Since the effect of DL on word order does not vary much by clause type (Fig. 2) and models that allow this variation leverage almost no weight (Table 3), we collapse over main and subordinate clauses in what follows.

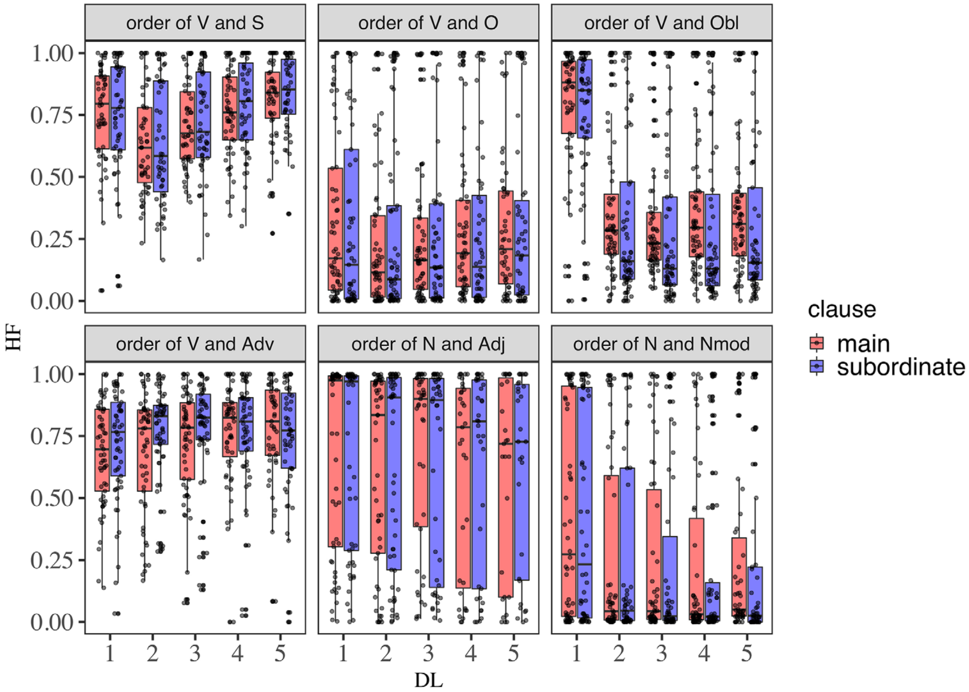


Fig. 2. Boxplots for the distribution of raw HF dependencies in relation to DL across main and subordinate clauses.

Table 3
Model comparisons via approximate leave-one-out cross-validation and model stacking. Models are ordered by their expected log pointwise densities (*elpd*) as estimated via Pareto-smoothed importance sampling (*psis*). A higher elpd value indicates better model performance

Model	$elpd_{psis}$	$\Delta elpd_{psis}$	$SE(\Delta elpd_{psis})$	Weights
Dependency type by clause type	−13,029.31	0.00	0.00	0.38
Dependency type	−13,049.78	−20.47	8.42	0.46
DL by dependency type	−14,059.59	−1,030.29	54.02	0.15
DL by clause type	−15,412.45	−2,383.14	53.7	0.01
DL	−15,416.04	−2,386.74	53.56	0.00
Clause type	−15,450.81	−2,421.5	51.94	0.00
Intercept	−15,457.31	−2,428.00	51.64	0.00

With regard to preferred head orderings, we observe strong differences between dependency types: subjects (“S”), adverbs (“Adv”), and adjectives (“Adj”) slightly prefer to precede their heads (i.e., have higher (log) odds for HF dependencies), while the other dependencies, especially objects and nominal modifiers, tend to follow their heads in our sample languages. However, with the exception of the initial placement of subjects, there is substantial cross-linguistic variation (as shown by the thin gray lines in the figure).

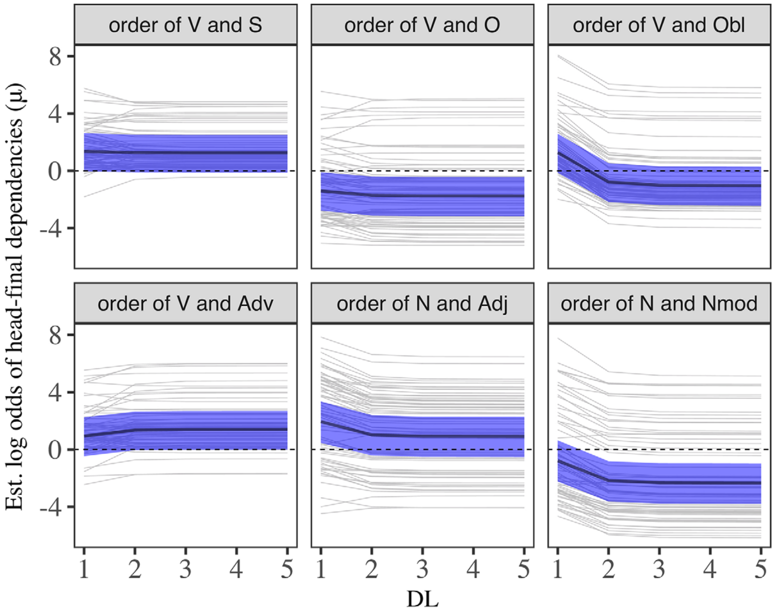


Fig. 3. Monotonic effect of DL on word order directions (μ in log odds) across dependency types, after conditioning on the language-specific effects ("random slopes") and the overall variation between families and areas ("random intercepts"). Higher μ values indicate higher (log) odds of HF orders. The thick black lines show the estimated mean effect of DL at the population level ("fixed effects"), and the thin gray lines show the estimated mean effects in individual languages ("random effects"). The colored ribbons indicate 90% credible intervals around the population-level mean.

The overall effect of DL on word order direction includes 0 (90% CI = $[-0.57, 0.23]$; $p(slope < 0) = .82$), but there is appreciable variation across dependency types: obliques (estimated mean slope: -0.56; 90% CI = $[-0.9, -0.2]$; $p(slope < 0) = .97$) and nominal modifiers (estimated mean slope: -0.38; 90% CI = $[-0.73, -0.03]$; $p(slope < 0) = .95$) show negative correlations, suggesting strong tendencies for long obliques and nominal modifiers to be post-posed (see Table 4 for summaries). The major differences in head directionality (on average 89%) occur between short dependencies at the lengths of 1 and 2 (cf. the simplex estimates of the monotonic parameterization of DL in Section S4.2 in the Supporting Information). By contrast, there are no strong correlations between DL and head-finality for the other orders, as all 90% CI includes 0. With the increase of DL, adjective modifiers show a slight negative trend (estimated mean slope: -0.25; 90% CI = $[-0.59, 0.1]$; $p(slope < 0) = .91$), whereas adverb modifiers exhibit a slight positive correlation (estimated mean slope: 0.13; 90% CI = $[-0.21, 0.47]$; $p(slope > 0) = .86$). The placements of subjects and objects seem to be independent of DL (see Table 4 for estimated slopes of DL on word order directions). The relationship between DL and head directionality is relatively consistent between population ("fixed effects") and language ("random effects") levels for obliques (range of estimated slopes: $[-0.83, -0.1]$; $SD = 0.16$), and nominal modifiers (range of estimated slopes: $[-0.65, 0.08]$; $SD = 0.16$). More language-specific relationships of DL on head ordering are

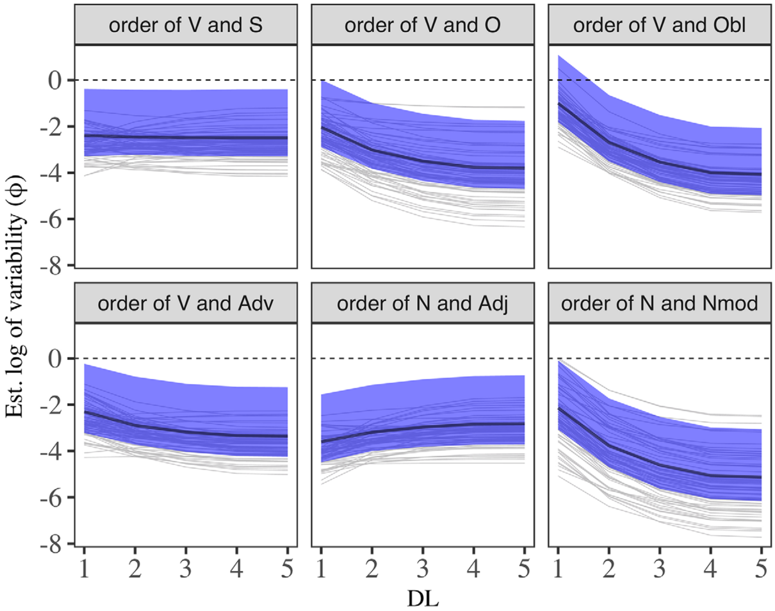


Fig. 4. Monotonic effect of DL on word order variation (ϕ on the log scale) across dependency types, after conditioning on the language-specific effects ("random slopes") and the overall variation between families and areas variable ("random intercepts") by family and area. Higher ϕ values indicate more word order variation. The thick black lines show the estimated mean effects of DL at the population level ("fixed effects"), and the thin gray lines show the estimated mean effects at the language level ("random effects"). The colored ribbons indicate 90% credible intervals around the population-level mean.

Table 4
Estimated intercepts and mean slopes for DL on word order directions (μ in log odds) across dependency types. The last column reports the proportion of the posterior estimates that is higher than 0 when the mean is higher than 0, and the proportion that is lower than 0 if the mean is lower than 0

Dependency Type	Parameter	Mean	90% CI	$p(parameter <\|> 0)$
V and S	Intercept	1.33	[−0.04, 2.63]	.95
	Slope	−0.02	[−0.37, 0.33]	.60
V and O	Intercept	−1.42	[−2.77, 0.10]	.96
	Slope	−0.08	[−0.41, 0.27]	.78
V and Obl	Intercept	1.25	[−0.13, 2.58]	.93
	Slope	−0.56	[−0.90, 0.22]	.97
V and Adv	Intercept	0.92	[−0.46, 2.25]	.87
	Slope	0.13	[−0.21, 0.47]	.86
N and Adj	Intercept	1.93	[0.46, 3.35]	.98
	Slope	−0.25	[−0.59, 0.10]	.91
N and Nmod	Intercept	−0.79	[−2.23, 0.63]	.83
	Slope	−0.38	[−0.73, −0.03]	.95

Table 5
Estimated intercepts and mean slopes for DL on word order variation (ϕ on the log scale) across dependency types. The last column reports the proportion of the posterior estimates that is higher than 0 when the mean is higher than 0, and the proportion that is lower than 0 if the mean is lower than 0

Dependency Type	Parameter	Estimates	90% CI	$p(parameter <\|> 0)$
V and S	Intercept	−2.21	[−3.28, −0.38]	.97
	Slope	0.01	[−0.39, 0.47]	.48
V and O	Intercept	−1.83	[−2.89, 0.01]	.95
	Slope	−0.42	[−0.79, 0.05]	.94
V and Obl	Intercept	−0.77	[−1.81, 1.08]	.84
	Slope	−0.75	[−1.13, −0.30]	.97
V and Adv	Intercept	−2.10	[−3.22, −0.23]	.96
	Slope	−0.24	[−0.62, 0.22]	.89
N and Adj	Intercept	−3.40	[−4.48, −1.56]	1.00
	Slope	0.21	[−0.17, 0.68]	.88
N and Nmod	Intercept	−1.94	[−3.07, −0.09]	.96
	Slope	−0.73	[−1.13, −0.28]	.97

observed for the other orders (cf. the thin gray lines in Fig. 3 and more detailed reports in Figs. S7– S12 and Tables S3– S8 in the Supporting Information).

Turning to the predictive effects of DL on word order variation in Fig. 4 and Table 5, we note similar extents of variation across all dependency types (with a mean value of about $\log(\phi) = -2.5$). There is no evidence for a general effect of DL across all types at the population level (90% CI = $[-0.83, 0.25]$; $p(slope < 0)$= .85) and effects differ across types: there are strong effects of DL on variation in obliques (estimated mean slope = −0.75; 90% CI = $[-1.13, -0.3]$; $p(slope < 0) = .97$) and nominal modifiers (estimated mean slope =−0.73; 90% CI = $[-1.13, -0.28]$; $p(slope < 0) = .97$), that is longer dependencies of obliques and nominal modifiers yield less word order variation. A strong negative effect of DL is observed for objects (estimated mean slope = −0.42; 90% CI = $[-0.79, 0.05]$; $p(slope < 0) = .94$) and adverbial modifiers (estimated mean slope = −0.24; 90% CI = $[-0.62, 0.22]$; $p(slope < 0) = .89$). A slight positive relationship between DL and variation is found for adjectives (estimated mean slope = 0.21; 90% CI = $[-0.17, 0.68]$; $p(slope > 0)$= .88), and no correlation is observed for subjects (estimated mean slope = 0.01; 90% CI = $[-0.39, 0.47]$; $p(slope > 0)$= .48). All the observations are remarkably consistent across individual languages. This can be seen by the consistent downward slopes of gray lines, which represent the individual languages in Fig. 4. The estimated slopes for obliques range from −1.13 to −0.41, with a standard deviation of 0.16, and the estimated slopes for nominal modifiers range from −1.11 to −0.39, with a standard deviation of 0.16 across different languages (see Figs. S14– S19 and Tables S10– S15 in the Supporting Information). The strongest effect occurs between lengths 1 and 2, although with 52% the difference to the other levels is not as big as in the case of word order direction (see Section S4.2).

4. Discussion

Our results suggest that DL is not sufficient to predict the directions and variation of word orders, and no general correlations are obtained across all dependency types. Strong effects of DL on direction and variation are limited to obliques and nominal modifiers: These dependency types show strong evidence for longer dependents to be placed after the head and more rigidly so than shorter dependents. This limitation of effects challenges across-the-board principles like Behaghel's law, the easy first principle or heaviness serialization principle, which all favor a consistently rightward positioning of longer dependents (or constituents) (Arnold et al., 2000; Behaghel, 1909; Hawkins, 1983; MacDonald, 2013; Wasow, 1997). Our results are also difficult to reconcile with the theory of predictability maximization, which would favor a consistent placement of dependents before the head across types (Ferrer-i-Cancho, 2017; Konieczny, 2000; Vasishth & Lewis, 2006).

While we did not formally test this, our findings are only partially consistent with the observation in DLM research that longer elements tend to be placed later in HI languages and earlier in HF languages, that is a language-level effect of harmonic ordering across types (Dryer, 1992; Futrell et al., 2020; Greenberg, 1963; Hawkins, 1983). This expectation shows up in our results in a partial segregation of language-specific estimates in verb/object and verb/subject dependencies, with some languages showing an increase of HF positions and others a decrease with length (also see Figs. S7 and S8 in the Supporting Information for more detailed plots). However, this segregation into HF versus HI types is weak since many languages fall in between, and slopes are inconsistent even in the more clearly HF (above 0) and more clearly HI (below 0) languages. This suggests that the expected language-level effect has only limited scope and that our results are not an artifact of pooling HI and HF languages with opposite length effects.

DL has a very weak effect on the order of adjectives and adverbs. While there is a general preference for placing them before the head (means are above 0 in Fig. 3) and for constraining their order (all values are below 0 in Fig. 4), the preferences go in opposite, albeit very weak directions: the order preference decreases and variation increases with DL in the case of adjectives; the order preference increases and variation decreases with DL in the case of adverbs. While the pattern we observed for adverbs is consistent with the notion that single-word dependents are more variably placed at the opposite side of the head to achieve DLM (Gulordava, 2018; Temperley, 2008; Temperley & Gildea, 2017), this does not generalize to adjectives.

Intriguingly, the effect of DL is not evenly distributed over lengths. The effects are mostly due to differences between adjacent and non-adjacent dependencies: on average 89% of the differences in direction, and 52% of the difference in variation are located between the first two levels of DL (lengths of 1 and 2). This implies that the increase of DL has a stronger negative effect for short dependencies, but has less influence on longer dependencies. Furthermore, it suggests a major distinction between single-word versus phrasal dependents. This is consistent with proposals that draw a categorical difference between non-branching and branching elements (Dryer, 1992), but the effect is limited to obliques and nominal modifiers.

Word order directions differ considerably across dependency types in Fig. 3, although they all show a consistent overall bias against word order variation 4. A strong and cross-linguistically robust preference for HF order is limited to the subject-verb (SV) dependency. This is likely to reflect an oft-noted preference to place agents and topics first (Bornkessel-Schlesewsky & Schlesewsky, 2009; Comrie, 1981; Gibson et al., 2013, 2019; Givón, 1983; Goldin-Meadow, So, Özyürek, & Mylander, 2008; Hall et al., 2013; Kemmerer, 2012; Krebs, Malaia, Wilbur, & Roehm, 2018; Meir, Sandler, Padden, & Aronoff, 2010; Riesberg, Malcher, & Himmelmann, 2019; Schouwstra & de Swart, 2014; Tomlin, 1986).

We also note a general bias towards HI VO order, although it varies strongly across individual languages. As noted above, with the increase of DL, we furthermore observe a somewhat divergent trend for initial or final placement of long dependents at the language level, and different languages seem to explore their own ways of ordering complex elements. Further research is needed to evaluate to what extent DL effects in an individual object/verb dependency interact with other dependencies in the same sentence or in the same language.

Assuming the VO bias holds independently of such potential interactions, it combines with the SV bias in a weak overall preference for subject-verb-object (SVO) order that emerges when estimates are conditioned on cross-linguistic variation in our sample. To some extent this finding might be driven by the overrepresentation of European languages in the sample although our models condition of areal and family relations. To the extent it holds, such a preference is consistent with DLM observations: when there are two simultaneous dependents (S and O) of the same head (V), the total DL is minimal when they occur on either side of the head (Ferrer-i Cancho, 2008, 2015). However, DLM observations elsewhere furthermore suggest that long dependents should be more constrained than short dependents. Yet this is not supported by our findings (Fig. 3): the directions of S and O barely change with increasing lengths.

Our results show very similar word order patterns across main and subordinate clauses. This is in line with previous findings of roughly similar amount of word order variation across contexts (Kroch, 1989, 2001), and it challenges theories that assume more variation in main than in subordinate clauses (Bybee, 2002; Givón, 1979; Hooper & Thompson, 1973; Ross, 1973) and therefore a higher innovation potential of main clauses (Bean, 1983; MacLeish, 1969; Stockwell & Minkova, 1991).

Previous work has also examined the relationship between word order direction and variation (Levshina, 2019). Our results are only partly consistent with the results from this work, although a full comparison is limited by the fact that we select narrower dependency types by relying on both dependency (“nsubj,” “obj,” “obl,” “advmod,” “amod,” and “nmod”) and part-of-speech tags (NOUN, VERB, ADJ, and ADV), whereas Levshina (2019) collapses some types, relies on dependency tags only, and examines variability depending on overall language type (VO vs. OV) rather than individual dependencies. Fig. 5 plots our estimates of variation against word order for dependency type. Overall, the correlations are relatively weak and virtually absent in the subject (S/V) and adverb (V/Adv) dependencies. One potential correlation seems to be a slightly lower variation in pre-nominal (log odds of HF > 0 in 5) than in postnominal (log odds < 0) adjectives (Greenberg, 1963). This is partially consistent with Levshina’s findings although her analysis combines adjectival with nominal modifiers,

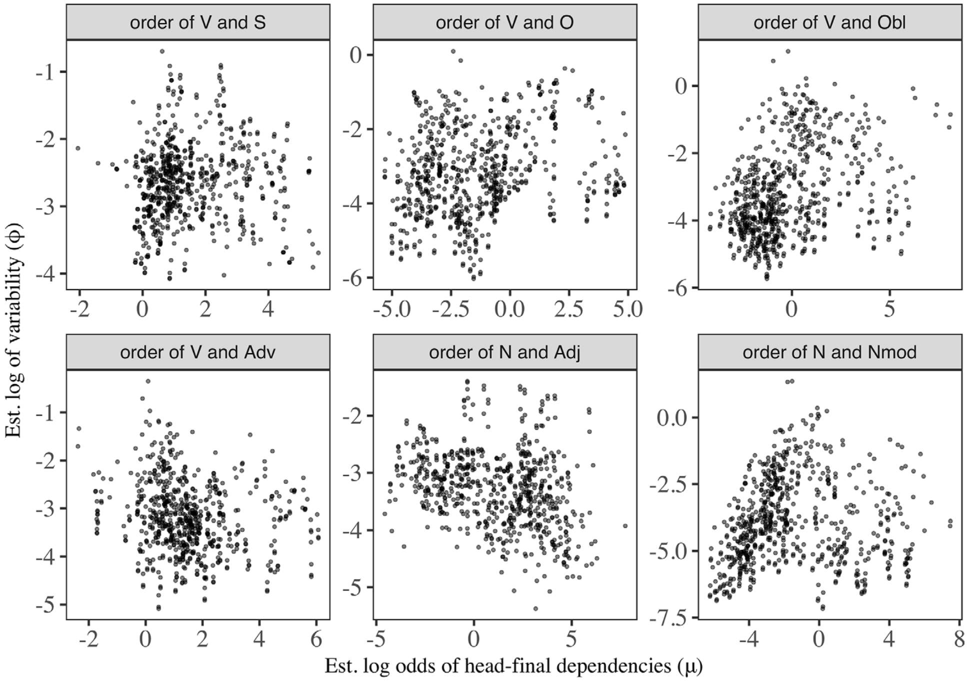


Fig. 5. Relationship between word order direction (μ) and variation (ϕ). The x-axis is the estimated log odds of HF dependencies, and the y-axis is the estimated log of variation. Each data point indicates the estimated direction and variation at a given combination of levels (DL, clause type, and dependency type), conditioned on the variation by language, language family, and area.

for which we found the opposite trend, that is a slight increase of variation in more HF dependencies. A possible reason for this discrepancy is that Levshina's analysis also includes numeral modifiers in the same overall noun-headed dependency type, and to the extent that numerals behave like adjectives, they might have strengthened the correlation despite the opposite trend among nominal modifiers.

For objects and obliques, we observe potentially more variation before the head (log odds of HF > 0) than after the head (Dryer, 1996; Dryer & Gensler, 2005; Hawkins, 2014). To the extent that these orders (unlike adjectives and adverbs) follow a general VO versus OV distinction at the language level (Dryer, 1992), our findings do not support Levshina's generalization of more variation among core arguments and oblique/adverbial dependencies in HI (VO) and flexible languages (i.e., log odds ≤ 0 in our terms) than in OV languages (log odds > 0). At first sight, a possible cause for this discrepancy might be subordinate clauses, which are included in Levshina's generalization but not in our study. However, the total number of datapoints that this inclusion would add is unlikely to be so large that they would not only fully reverse the trend we see in 5 but also strengthen it at the same time. Another possibility is that Levshina includes S/V and Adv/V orders among core arguments and oblique/adverbial dependencies, respectively, but our findings on these suggest that they would at best decrease the strength of the positive trend we find, but again, not fully reverse and at the same time sharpen it.

In view of this, we suspect that the most likely reason for the discrepancy across all dependencies is that entropy measures might have increased error rates in sparse regions (Chao et al., 2013; DeDeo et al., 2013) and that language-level estimates of word order types might miss finer distinctions between individual dependencies and their ordering. More methodological research, ideally with simulation studies, is needed to fully resolve the issue.

5. Conclusions

Unlike previous work, we estimated word order and word order variation at the level of individual dependencies. This revealed substantially more diversity than current theories would predict. Our results support the general expectation of DL effects on word order only for two out of six types of dependencies: obliques (e.g., “over the weekend” in “read the book over the weekend”) and nominal modifiers (e.g., “of the cabin” in “the owner of the cabin”) show an increase of final and more rigid placement, with the difference being driven mostly by the distinction between single-word and multiword dependents. While this is consistent with theories of DLM and theories that favor early placement of simple items, all these theories overgenerate, since they make the same predictions for other dependency types as well. Yet we found only very partial or no effects in the other types: Objects and adverbs show an effect for word order variation, but not head placement; adjectives show a very weak effect for head placement but an (equally weak) inverse effect for word order variation; subjects show no effect for either measure.

Overall then, this suggests limitations of current theories and invites more detailed experimental and corpus-based research on why DL effects are limited to certain dependency types. Given that the dependency types that we conditioned on rest on a combination of syntactic with semantic criteria (de Marneffe & Manning, 2008; Nivre, 2015), we propose that future research might profit from relativizing expected effects to specific kinds and degrees of dependencies (e.g. in terms of mutual information). At the same time, future models need to take into account the major difference we found between single-word and multiword dependencies, and they might need to further control for interaction effects between dependency types at the level of individual sentences or entire languages.

This invites a reconceptualization of research on how syntactic complexity interacts with order and, more specifically of research on DLM. Rather than assessing effects at the level of entire grammars, we submit that future theories need to carefully disentangle effects from individual dependencies and effects from the interaction between dependencies. In other words, progress might need a similar move from holistic grammars to specific properties and their interactions that have characterized progress in linguistic typology (Bickel, 2007). As our results suggest, holistic research might miss important differences between dependencies.

Acknowledgments

This research was partly supported by the Ph.D grant from the China Scholarship Council (201606320224) and the Postdoc mobility grant from the Swiss National Science Foundation (P500PH_202882) to Yingqi Jing. We would like to thank the anonymous reviewers and our colleagues, in particular, Lena A. Jäger and Paola Merlo, for helpful comments on an earlier draft of the paper.

Open access funding provided by Universitat Zurich.

Open Research Badges

This article has earned Open Data and Open Materials badges. Data are available at https://universaldependencies.org/ and materials are available at https://osf.io/4s67a/.

Notes

1. https://universaldependencies.org/.
2. Ideally, one would want to extract the effect estimates from the stacked model, but this is currently not possible for other response terms than the mean in a distributional model. The full distributional model is formulated as follows (again using `brms` notation): $\mu \sim \mathrm{DL} + \mathrm{CLS} + \mathrm{DEP} + (0 + \mathrm{DEP} + \mathrm{DL} \mid \mathrm{CLS}) + (0 + \mathrm{DL} \mid \mathrm{DEP}) + (1 + \mathrm{CLS} + \mathrm{DEP} + \mathrm{DL} \mid \mathrm{lang}) + (1 \mid \mathrm{fam}) + (1 \mid \mathrm{area})$, and $\phi \sim \mathrm{DL} + \mathrm{CLS} + \mathrm{DEP} + (0 + \mathrm{DEP} + \mathrm{DL} \mid \mathrm{CLS}) + (0 + \mathrm{DL} \mid \mathrm{DEP}) + (1 + \mathrm{CLS} + \mathrm{DEP} + \mathrm{DL} \mid \mathrm{lang}) + (1 \mid \mathrm{fam}) + (1 \mid \mathrm{area})$.

References

Arnold, J. E., Losongco, A., Wasow, T., & Ginstrom, R. (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, *76*(1), 28–55. https://doi.org/10.2307/417392

Bean, M. C. (1983). *The development of word order patterns in Old English*. London: Croom Helm.

Behaghel, O. (1909). Beziehungen zwischen umfang und reihenfolge von satzgliedern. *Indogermanische Forschungen*, *25*, 110–142. https://doi.org/10.1515/9783110242652.110

Bickel, B. (2007). Typology in the 21st century: Major current developments. *Linguistic Typology*, *11*, 239–251. https://doi.org/10.1515/LINGTY.2007.018

Bock, J. K. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review*, *89*(1), 1–47. https://doi.org/10.1037/0033-295X.89.1.1

Bock, J. K., & Levelt, W. (1994). Language production: Grammatical encoding. In M. Gernsbacher (Ed.) *Handbook of psycholinguistics* (pp. 945–984). New York, NY: Academic Press.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2009). The role of prominence information in the real-time comprehension of transitive constructions: A cross-linguistic approach. *Language and Linguistics Compass*, *3*(1), 19–58. https://doi.org/10.1111/j.1749-818X.2008.00099.x

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Define custom response distributions with brms. Talk presented at the StanCon 2018. Helsinki, Finland.

Bürkner, P.-C., & Charpentier, E. (2020). Modeling monotonic effects of ordinal predictors in Bayesian regression models. *British Journal of Mathematical and Statistical Psychology* (pp. 420–451). https://doi.org/10.1111/bmsp.12195

Bybee, J. (2002). Main clauses are innovative, subordinate clauses are conservative: Consequences for the nature of constructions. In J. Bybee, M. Noonan, & S. Thompson (Eds.) *Complex sentences in grammar and discourse: Essays in honor of Sandra A. Thompson* (pp. 1–17). Amsterdam, The Netherlands: John Benjamins.

Chao, A., Wang, Y. T., & Jost, L. (2013). Entropy and the species accumulation curve: A novel entropy estimator via discovery rates of new species. *Methods in Ecology and Evolution*, *4*(11), 1091–1100. https://doi.org/10.1111/2041-210X.12108

Choi, H.-W. (2007). Length and order: A corpus study of Korean dative-accusative constructions. *Discourse and Cognition*, *14*, 207–227. https://doi.org/10.15718/DISCOG.2007.14.3.207

Christianson, K., & Ferreira, F. (2005). Conceptual accessibility and sentence production in a free word order language (Odawa). *Cognition*, *98*(2), 105–135. https://doi.org/10.1016/j.cognition.2004.10.006

Comrie, B. (1981). *Language universals and linguistic typology*. Oxford, England: Basil Blackwell.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, *122*(3), 306–329. https://doi.org/10.1016/j.cognition.2011.10.017

de Marneffe, M.-C., & Manning, C. D. (2008). The Stanford typed dependencies representation. In *Proceedings of the Workshop on Cross-framework and Cross-domain Parser Evaluation* (pp. 1–8). Stroudsburg, PA: Association for Computational Linguistics.

de Marneffe, M.-C., & Nivre, J. (2019). Dependency grammar. *Annual Review of Linguistics*, *5*, 197–218. https://doi.org/10.1146/annurev-linguistics-011718-011842

DeDeo, S., Hawkins, R. X., Klingenstein, S., & Hitchcock, T. (2013). Bootstrap methods for the empirical study of decision-making and information flows in social systems. *Entropy*, *15*(6), 2246–2276. https://doi.org/10.3390/e15062246

Dryer, M. S. (1992). The Greenbergian word order correlations. *Language*, *68*(1), 81–138. https://doi.org/10.2307/416370

Dryer, M. S. (1996). Word order typology. In J. Jacobs (Ed.) *Handbook on syntax* (Vol. *2*, pp. 1050–1065). Berlin, Germany: Walter de Gruyter.

Dryer, M. S., & Gensler, O. D. (2005). Order of object, oblique, and verb. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online* (pp. 330–341). Leipzig, Germany: Max Planck Institute for Evolutionary Anthropology. https://wals.info/chapter/84

Faghiri, P., & Samvelian, P. (2014). Constituent ordering in Persian and the weight factor. In P. Christopher (Ed.) *Empirical issues in syntax and semantics 10* (pp. 215–232). Paris, France: CNRS.

Faghiri, P., & Samvelian, P. (2020). Word order preferences and the effect of phrasal length in SOV languages: Evidence from sentence production in persian. *Glossa: A Journal of General Linguistics*, *5*(1), 1–33. https://doi.org/10.5334/gjgl.1078

Fedzechkina, M., Chu, B., & Florian Jaeger, T. (2018). Human information processing shapes language change. *Psychological Science*, *29*(1), 72–82. https://doi.org/10.1177/0956797617728726

Ferrer i Cancho, R. (2004). Euclidean distance between syntactically linked words. *Physical Review E*, *70*(5), 056135. https://doi.org/10.1103/PhysRevE.70.056135

Ferrer-i Cancho, R. (2008). Some word order biases from limited brain resources: A mathematical approach. *Advances in Complex Systems*, *11*(3), 393–414. https://doi.org/10.1142/S0219525908001702

Ferrer-i Cancho, R. (2015). The placement of the head that minimizes online memory: A complex systems approach. *Language Dynamics and Change*, *5*(1), 114–137. https://doi.org/10.1163/22105832-00501007

Ferrer-i-Cancho, R. (2017). The placement of the head that maximizes predictability. An information theoretic approach. *Glottometrics*, *39*, 38–71.

Futrell, R., Levy, R., & Gibson, E. (2020). Dependency locality as an explanatory principle for word order. *Language*, *96*(2), 371–412. https://doi.org/10.1353/lan.2020.0024

Futrell, R., Mahowald, K., & Gibson, E. (2015). Quantifying word order freedom in dependency corpora. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)* (pp. 91–100). Uppsala, Sweden: Uppsala University.

Gabry, J., & Češnovar, R. (2020). *cmdstanr: R interface to CmdStan*. https://mc-stan.org/cmdstanr

Gennari, S. P., Sloman, S. A., Malt, B. C., & Fitch, W. T. (2002). Motion events in language and cognition. *Cognition*, *83*, 49–49. https://doi.org/10.1016/S0010-0277(01)00166-4

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76. https://doi.org/10.1016/S0010-0277(98)00034-1

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain: Papers from the first mind articulation project symposium* (pp. 95–126). Cambridge, MA: MIT Press.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5), 389–407. https://doi.org/10.1016/j.tics.2019.02.003

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013). A noisy-channel account of crosslinguistic word-order variation. *Psychological Science*, *24*(7), 1079–1088. https://doi.org/10.1177/0956797612463705

Gildea, D., & Jaeger, T. F. (2015). *Human languages order information efficiently*. Preprint. Retrieved from arXiv:1510.02823.

Gildea, D., & Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science*, *34*(2), 286–310. https://doi.org/10.1111/j.1551-6709.2009.01073.x

Givón, T. (1979). *On understanding grammar*. New York, NY: Academic Press.

Givón, T. (1983). Topic continuity and word order pragmatics in Ute. In T. Givón (Ed.), *Topic continuity in discourse. A quantitative cross-language study* (pp. 141–214). Amsterdam, The Netherlands: John Benjamins.

Givón, T. (1988). The pragmatics of word order: Predictability, importance and attention. In M. Hammond, E. A. Moravcsik, & W. Jessica (Eds.), *Studies in syntactic typology* (pp. 243–284). Amsterdam, The Netherlands: John Benjamins Publishing Company.

Goldin-Meadow, S., So, W. C., Özyürek, A., & Mylander, C. (2008). The natural order of events: How speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences*, *105*(27), 9163–9168. https://doi.org/10.1073/pnas.0710060105

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg (Ed.), *Universals of language* (pp. 58–60). Cambridge, MA: MIT Press.

Groß, T., & Osborne, T. (2015). The dependency status of function words: Auxiliaries. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)* (pp. 111–120). Uppsala, Sweden: Uppsala University.

Gulordava, K. (2018). *Word order variation and dependency length minimisation: a cross-linguistic computational approach* (Ph.D. thesis). University of Geneva, Geneva, Switzerland.

Gulordava, K., & Merlo, P. (2015). Diachronic trends in word order freedom and dependency length in dependency-annotated corpora of Latin and ancient Greek. In *Proceedings of the Third International conference on dependency linguistics (Depling 2015)* (pp. 121–130). Uppsala, Sweden: Uppsala University.

Gundel, J. K. (1988). Universals of topic-comment structure. In M. Hammond, E. A. Moravcsik, & W. Jessica (Eds.), *Studies in syntactic typology* (pp. 209–239). Amsterdam, The Netherlands: John Benjamins.

Gundel, J. M. (1975). *The role of topic and comment in linguistic theory*. Chur, Switzerland: Harwood Academic.

Hall, M. L., Mayberry, R. I., & Ferreira, V. S. (2013). Cognitive constraints on constituent order: Evidence from elicited pantomime. *Cognition*, *129*(1), 1–17. https://doi.org/10.1016/j.cognition.2013.05.004

Hausser, J., & Strimmer, K. (2009). Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, *10*(7), 1469–1484. Retrieved from https://www.jmlr.org/papers/volume10/hausser09a/hausser09a.pdf

Haviland, S. E., & Clark, H. H. (1974). What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior*, *13*(5), 512–521. https://doi.org/10.1016/S0022-5371(74)80003-4

Hawkins, J. A. (1983). *Word order universals*. New York, NY: Academic Press.

Hawkins, J. A. (1994). *A performance theory of order and constituency*. Cambridge, England: Cambridge University Press.

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

Hock, H. H. (1991). *Principles of historical linguistics*. Berlin, Germany: Walter de Gruyter.

Hooper, J. B., & Thompson, S. A. (1973). On the applicability of root transformations. *Linguistic Inquiry*, *4*(4), 465–497.

Hudson, R. A. (1984). *Word grammar*. Oxford, England: Blackwell Oxford.

Husain, S., Vasishth, S., & Srinivasan, N. (2014). Strong expectations cancel locality effects: Evidence from Hindi. *PloS ONE*, *9*(7), e100986. https://doi.org/10.1371/journal.pone.0100986

Jing, Y., Blasi, D. E., & Bickel, B. (2021). Dependency length minimization and its limits: a possible role for a probabilistic version of the Final-Over-Final Condition (to appear in Language). https://doi.org/10.31234/osf.io/sp7r2

Jørgensen, B. (1997). *The theory of dispersion models*. Boca Raton, FL: CRC Press.

Kemmerer, D. (2012). The cross-linguistic prevalence of SOV and SVO word orders reflects the sequential and hierarchical representation of action in Broca's area. *Language and Linguistics Compass*, *6*, 50–66. https://doi.org/10.1002/lnc3.322

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research*, *29*(6), 627–645. https://doi.org/10.1023/A:1026528912821

Krebs, J., Malaia, E., Wilbur, R. B., & Roehm, D. (2018). Subject preference emerges as cross-modal strategy for linguistic processing. *Brain Research*, *1691*, 105–117. https://doi.org/10.1016/j.brainres.2018.03.029

Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica*, *55*(3-4), 243–276. https://doi.org/10.1556/ALing.55.2008.3-4.2

Kroch, A. S. (1989). Reflexes of grammar in patterns of language change. *Language Variation and Change*, *1*(3), 199–244. https://doi.org/10.1017/S0954394500000168

Kroch, A. S. (2001). Syntactic change. In M. R. Baltin, & C. Collins (Eds.), *The handbook of contemporary syntactic theory* (pp. 699–729). Malden, MA: Blackwell.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. New York, NY: Academic Press.

Lehmann, W. P. (1973). A structural principle of language and its implications. *Language* (pp. 47–66). https://doi.org/10.2307/412102

Levelt, W. J. (1999). Producing spoken language: A blueprint of the speaker. In C. Brown, & P. Hagoort (Eds.), *The neurocognition of language* (pp. 83–122). Oxford, England: Oxford University Press.

Levshina, N. (2019). Token-based typology and word order entropy: A study based on universal dependencies. *Linguistic Typology*, *23*(3), 533–572. https://doi.org/10.1515/lingty-2019-0025

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Levy, R. P., & Keller, F. (2013). Expectation and locality effects in German verb-final structures. *Journal of Memory and Language*, *68*(2), 199–222. https://doi.org/10.1016/j.jml.2012.02.005

Lightfoot, D. (1982). *The language lottery: Toward a biology of grammars*. Cambridge, MA: MIT Press.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, *9*(2), 159–191. https://doi.org/10.17791/JCS.2008.9.2.159

Liu, H. (2009). *Dependency Grammar from theory to practice*. Beijing, China: Science Press.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*, 226. https://doi.org/10.3389/fpsyg.2013.00226

MacLeish, A. (1969). *The Middle English subject-verb cluster*. The Hague, The Netherlands: Mouton.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton, FL: CRC press.

Meir, I., Sandler, W., Padden, C., & Aronoff, M. (2010). Emerging sign languages. In M. Marschark & P. E. Spencer (Eds.), *Oxford handbook of deaf studies, language, and education* (Vol. *2*, pp. 267–280). Oxford, England: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195390032.013.0018

Melčuk, I. (1988). *Dependency syntax: Theory and practice*. Albany, NY: State University Press of New York.

Nalborczyk, L., Batailler, C., Lœvenbruck, H., Vilain, A., & Bürkner, P.-C. (2019). An introduction to Bayesian multilevel models using brms: A case study of gender effects on vowel variability in standard Indonesian. *Journal of Speech, Language, and Hearing Research*, *62*(5), 1225–1242. https://doi.org/10.1044/2018_JSLHR-S-18-0006

Nemenman, I., Shafee, F., & Bialek, W. (2002). Entropy and inference, revisited. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems* Vol. *14* (pp. 471–478). Cambridge, MA: MIT Press.

Nivre, J. (2015). Towards a universal grammar for natural language processing. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing* (pp. 3–16). Lecture Notes in Computer Science, Vol *9041*. Cham, Switzerland: Springer International.

Nivre, J., Abrams, M., Agić, Ž., Ahrenberg, L., Aleksandravičiūtė, G., Antonsen, L., Aplonova, K., Aranzabe, M. J., Arutie, G. & Asahara, M. et al. (2020). Universal dependencies 2.7. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University. Retrieved from http://hdl.handle.net/11234/1-3105

Osborne, T., & Gerdes, K. (2019). The status of function words in dependency grammar: A critique of universal dependencies (UD). *Glossa*, *4*(1), 1–28. https://doi.org/10.5334/gjgl.537

Osborne, T., & Maxwell, D. (2015). A historical overview of the status of function words in dependency grammar. *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)* (pp. 241–250). Uppsala, Sweden: Uppsala University.

Petrov, S., Das, D., & McDonald, R. (2012). A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 2089–2096). Luxembourg: European Language Resources Association.

Quirk, R. (1972). *A grammar of contemporary English*. Harlow, England: Longman Group.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Boston, MA: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Riesberg, S., Malcher, K., & Himmelmann, N. P. (2019). How universal is agent-first? evidence from symmetrical voice languages. *Language*, *95*(3), 523–561. https://doi.org/10.1353/lan.2019.0055

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

Ross, J. R. (1973). The penthouse principle and the order of constituents. In C. W. Corum, T. C. Smith-Stark, & A. Weiser (Eds.) *You take the high node and I'll take the low node* (pp. 397–422). Chicago Linguistic Society.

Sauppe, S., Choudhary, K. K., Giroud, N., Blasi, D. E., Norcliffe, E., Bhattamishra, S., Gulati, M., Egurtzegi, A., Bornkessel-Schlesewsky, I., Meyer, M., & Bickel, B. (2021). Neural signatures of syntactic variation in speech planning. *PLoS Biology*, *19*(1), e3001038. https://doi.org/10.1371/journal.pbio.3001038

Schouwstra, M., & de Swart, H. (2014). The semantic origins of word order. *Cognition*, *131*(3), 431–436. https://doi.org/10.1016/j.cognition.2014.03.004

Stan Development Team (2020). RStan: The R interface to Stan. Retrieved from http://mc-stan.org/

Stockwell, R., & Minkova, D. (1991). Subordination and word order change in the history of English. In D. Kastovsky (Ed.), *Historical English syntax* (pp. 367–408). Berlin, Germany: Walter de Gruyter.

Temperley, D. (2008). Dependency-length minimization in natural and artificial languages. *Journal of Quantitative Linguistics*, *15*(3), 256–282. https://doi.org/10.1080/09296170802159512

Temperley, D., & Gildea, D. (2017). Minimizing syntactic dependency lengths: Typological/cognitive universal? *Annual Review of Linguistics*, *4*(1), 1–15. https://doi.org/10.1146/annurev-linguistics-011817-045617

Tily, H. J. (2010). *The role of processing complexity in word order variation and change* (Ph.D. thesis). Stanford University, Stanford, CA.

Tomlin, R. (1986). *Basic word order: Functional principles*. London: Croom Helm.

Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, *82*, 767–794. https://doi.org/10.1353/lan.2006.0236

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2020). loo: Efficient leave-one-out cross-validation and waic for Bayesian models. Retrieved from https://mc-stan.org/loo

Vennemann, T. (1974). Topics, subjects, and word order: from SXV to SVX via TVX. In J. Anderson & C. Jones (Eds.), *Historical linguistics* (pp. 339–376). Amsterdam, The Netherlands: North Holland.

Vennemann, T. (1975). An explanation of drift. In C. N. Li (Ed.), *Word order and word order change* (pp. 267–305). Austin: University of Texas Press.

Wasow, T. (1997). End-weight from the speaker's perspective. *Journal of Psycholinguistic Research*, *26*(3), 347–361. https://doi.org/10.1023/A:1025080709112

Wolpert, D. H., & Wolf, D. R. (1995). Estimating functions of probability distributions from a finite set of samples. *Physical Review E*, *52*(6), 6841. https://doi.org/10.1103/PHYSREVE.52.6841

Yamashita, H. (2002). Scrambled sentences in japanese: Linguistic properties and motivations for production. *Text*, *22*(4), 597–634. https://doi.org/10.1515/text.2002.023

Yamashita, H., & Chang, F. (2001). "Long before short" preference in the production of a head-final language. *Cognition*, *81*, B45–B55. https://doi.org/10.1016/s0010-0277(01)00121-4

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, *13*(3), 917–1007. https://doi.org/10.1214/17-BA1091

Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (pp. 28–30). Luxembourg: European Language Resources Association.