Quantifying Word Order Freedom in Dependency Corpora

Richard Futrell, Kyle Mahowald, and Edward Gibson
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
{futrell, kylemaho, egibson}@mit.edu

Abstract

Using recently available dependency corpora, we present novel measures of a key quantitative property of language, *word order freedom*: the extent to which word order in a sentence is free to vary while conveying the same meaning. We discuss two topics. First, we discuss linguistic and statistical issues associated with our measures and with the annotation styles of available corpora. We find that we can measure reliable upper bounds on word order freedom in head direction and the ordering of certain sisters, but that more general measures of word order freedom are not currently feasible. Second, we present results of our measures in 34 languages and demonstrate a correlation between quantitative word order freedom of subjects and objects and the presence of nominative-accusative case marking. To our knowledge this is the first large-scale quantitative test of the hypothesis that languages with more word order freedom have more case marking (Sapir, 1921; Kiparsky, 1997).

1 Introduction

Comparative cross-linguistic research on the quantitative properties of natural languages has typically focused on measures that can be extracted from unannotated or shallowly annotated text. For example, probably the most intensively studied quantitative properties of language are Zipf's findings about the power law distribution of word frequencies (Zipf, 1949). However, the properties of languages that can be quantified from raw text are relatively shallow, and are not straightforwardly related to higher-level properties of languages such as their morphology and syntax. As a result, there has been relatively little large-scale comparative work on quantitative properties of natural language *syntax*.

In recent years it has become possible to bridge that gap thanks to the availability of large dependency treebanks for many languages and the development of standardized annotation schemes (de Marneffe et al., 2014; Nivre, 2015; Nivre et al., 2015). These resources make it possible to perform direct comparisons of quantitative properties of dependency trees. Some important previous work using dependency corpora to study crosslinguistic syntactic phenomena is Liu (2010), who quantifies the frequency of right- and left-branching in dependency corpora, and Kuhlmann (2013), who quantifies the frequency with which natural language dependency trees deviate from projectivity. Other work has studied graph-theoretic properties of dependency trees in the context of language classification (Liu and Li, 2010; Abramov and Mehler, 2011).

Here we study a particular quantitative property of language syntax: word order freedom. We focus on developing linguistically interpretable measures, as close as possible to an intuitive, relatively theory-neutral idea of what word order freedom means. In doing so, a number of methodological issues and questions arise. What quantitative measures map most cleanly onto the concept of word order freedom? Is it feasible to estimate the proposed measure given limited corpus size? Which corpus annotation style—e.g., content-head dependencies or dependencies where function words are heads—best facilitates crosslinguistic comparison? In this work, we argue for a set of methodological decisions which we believe balance the interests of linguistic interpretability, stability with respect to corpus size, and comparability across languages.

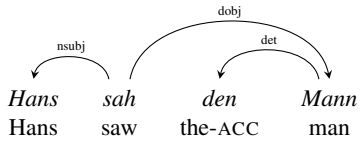
We also present results of our measures as applied to 34 languages and discuss their linguis-

tic significance. In particular, we find that languages with quantitatively large freedom in their ordering of subject and object all have nominative/accusative case marking, but that languages with such case marking do not necessarily have much word order freedom. This asymmetric relationship has been suggested in the typological literature (Kiparsky, 1997), but this is the first work to verify it quantitatively. We also discuss some of the exceptions to this generalization in the light of recent work on information-theoretic properties of different word orders (Gibson et al., 2013).

2 Word Order and the Notion of Dependency

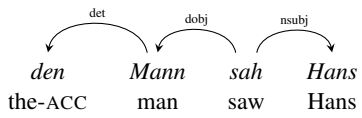
We define *word order freedom* as the extent to which the same word or constituent in the same form can appear in multiple positions while retaining the same propositional meaning and preserving grammaticality. For example, the sentence pair (1a-b) provides an example of word order freedom in German, while sentence pair (2a-b) provides an example of a lack of word order freedom in English. However, the sentences (2a) and (2c) do *not* provide an instance of word order freedom in English by our definition, since the agent and patient appear in different syntactic forms in (2c) compared to (2a). We provide dependency syntax analyses of these sentences below.

(1a)



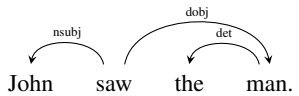
Meaning: "Hans saw the man."

(1b)

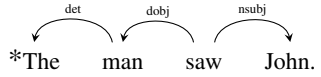


Meaning: "Hans saw the man."

(2a)

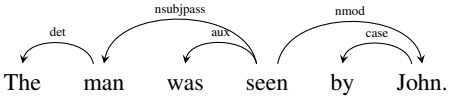


(2b)



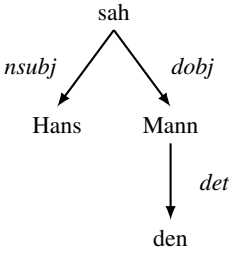
Cannot mean: "John saw the man."

(2c)



In the typological literature, this phenomenon has also been called *word order flexibility*, *pragmatic word order*, and a lack of *word order rigidity*. These last two terms reflect the fact that word order freedom does not mean that that word order is random. When word order is "free", speakers might order words to convey non-propositional aspects of their intent. For example, a speaker might place certain words earlier in a sentence in order to convey that those words refer to old information (Ferreira and Yoshita, 2003); a speaker might order words according to how accessible they are psycholinguistically (Chang, 2009); etc. Word order may be predictable given these goals, but here we are interested only in the extent to which word order is conditioned on the syntactic and compositional semantic properties of an utterance.

In a dependency grammar framework, we can conceptualize word order freedom as variability in the linear order of words given an unordered dependency graph with labelled edges. For example, both sentences (1a) and (1b) are linearizations of this unordered dependency graph:



The dependency formalism also gives us a framework for a functional perspective on why word order freedom exists and under what conditions it might arise. In general, the task of understanding the propositional meaning of a sentence requires identifying which words are linked to other words, and what the relation types of those links are. The dependency formalism directly encodes a subset of these links, with the additional assumption that links are always between exactly two explicit words. Therefore, we can roughly view an utterance as an attempt by a language producer to serialize a dependency graph such that a comprehender can recover it. The producer will want to choose a serialization which is efficient to

produce and which will allow the comprehender to recover the structure robustly. That is, the utterance must be informative about which pairs of words are linked in a dependency, and what the relation types of those links are.

Here we focus on the communication of relation types. In the English and German examples above, the relation types to be conveyed are *nsubj* and *dobj* in the notation of the Universal Dependencies project (Nivre et al., 2015). For the task of communicating the relation type between a head and dependent, natural languages seem to adopt two non-exclusive solutions: either the order of the head, the dependent, and the dependent's sisters is informative about relation type (a word order code), or the wordform of the head or dependent is informative about relation type (Nichols, 1986) (a case-marking code). Considerations of robustness and efficiency lead to a prediction of a tradeoff between these options. If a language uses case-marking to convey relation type, then word order can be repurposed to efficiently convey other, potentially non-propositional aspects of meaning. On the other hand, if a language uses inflexible word order to convey relation type, then it would be inefficient to also include case marking. However, some word order codes are less robust to noise than others (Gibson et al., 2013; Futrell et al., 2015), so certain rigid word orders might still require case-marking to maintain robustness. Similarly, some case-marking systems might be more or less robust, and so require rigid word order.

The idea that word order freedom is related to the prevalence of morphological marking is an old one (Sapir, 1921). A persistent generalization in the typological literature is that while word order freedom implies the existence of morphological marking, morphological marking does not imply the existence of word order freedom (Kiparsky, 1997; McFadden, 2003). These generalizations have been made primarily on the basis of native speaker intuitions and analyses of small datasets. Such data is problematic for measures such as word order freedom, since languages may vary quantitatively in how much variability they have, and it is not clear where to discretize this variability in order to form the categories "free word order" and "fixed word order". In order to test the reality of these generalizations, and to explore explanatory hypotheses for crosslinguistic variation, it is necessary to quantify the degree of word order freedom in a language.

3 Entropy Measures

Our basic idea is to measure the extent to which the linear order of words is determined by the unordered dependency graph of a sentence. A natural way to quantify this is *conditional entropy*:

$$H(X|C) = \sum_{c \in C} p_C(c) \sum_{x \in X} p_{X|C}(x|c) \log p_{X|C}(x|c), \quad (1)$$

which is the expected conditional uncertainty about a discrete random variable X, which we call the *dependent variable*, conditioned on another discrete random variable C, which we call the *conditioning variable*. In our case, the "perfect" measure of word order freedom would be the conditional entropy of sequences of words given unordered dependency graphs. Directly measuring this quantity is impractical for a number of reasons, so we will explore a number of entropy measures over partial information about dependency trees.

Using a conditional entropy measure with dependency corpora requires us to decide on three parameters: (1) the method of estimating entropy from observed joint counts of X and C, (2) the information contained in the dependent variable X, and (3) the information contained in the conditioning variable C. The two major factors in deciding these parameters are avoiding data sparsity and retaining linguistic interpretability. In this section we discuss the detailed considerations that must go into these decisions.

3.1 Estimating Entropy

The simplest way to estimate entropy given joint counts is through maximum likelihood estimation. However, maximum likelihood estimates of entropy are known to be biased and highly sensitive to sample size (Miller, 1955). The bias issues arise because the entropy of a distribution is highly sensitive to the shape of its tail, and it is difficult to estimate the tail of a distribution given a small sample size. As a result, entropy is systematically underestimated. These issues are exacerbated when applying entropy measures to natural language data, because of the especially long-tailed frequency distribution of sentences and words.

The bias issue is especially acute when doing crosslinguistic comparison with dependency cor-

pora because the corpora available vary hugely in their sample size, from 1017 sentences of Irish to 82,451 sentences of Czech. An entropy difference between one language and another might be the result of sample size differences, rather than a real linguistic difference.

We address this issue in two ways: first, we estimate entropy using the bootstrap estimator of DeDeo et al. (2013), and apply the estimator to equally sized subcorpora across languages[1]. Second, we choose dependent and conditioning variables to minimize data sparsity and avoid long tails. In particular, we avoid entropy measures where the conditioning variable involves wordforms or lemmas. We evaluate the effects of data sparsity on our measures in Section 4.

3.2 Local Subtrees

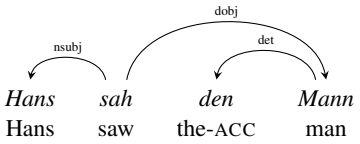
In order to cope with data sparsity and long-tailed distributions, the dependent and conditioning variables must have manageable numbers of possible values. This means that we cannot compute something like the entropy over full sentences given full dependency graphs, as these joint counts would be incredibly sparse, even if we include only part of speech information about words.

We suggest computing conditional entropy only on *local subtrees*: just subtrees consisting of a head and its immediate dependents. We conjecture that most word order and morphological rules can be stated in terms of heads and their dependents, or in terms of sisters of the same head. For example, almost all agreement phenomena in natural language involve heads and their immediate dependents (Corbett, 2006). Prominent and successful generative models of dependency structure such as the Dependency Model with Valence (Klein and Manning, 2004) assume that dependency trees are generated recursively by generating these local subtrees.

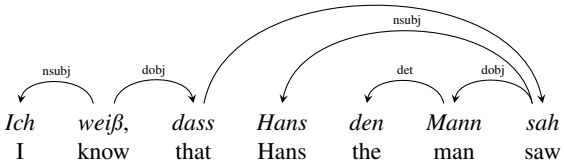
There are two shortcomings to working only with local subtrees; here we discuss how to deal with them.

First, there are certain word order phenomena which appear variable given only local subtree structure, but which are in fact deterministic given dependency structure beyond local subtrees. The extent to which this is true depends on the specifics of the dependency formalism. For example, in German, the position of the verb depends on clause type. In a subordinate clause with a complementizer, the verb must appear after all of its dependents (V-final order). Otherwise, the verb must appear after exactly one of its dependents (V2 order). If we analyze complementizers as heading their verbs, as in (3a), then the local subtree of the verb *sah* does not include information about whether the verb is in a subordinate clause or not.

(3a)

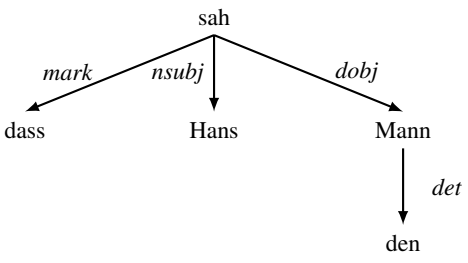


(3b)

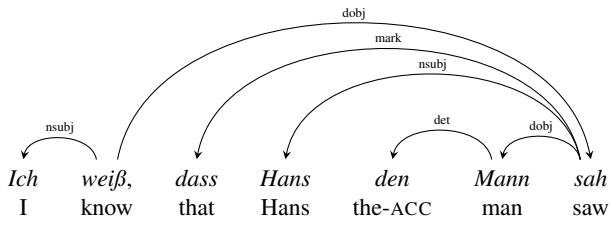


As a result, if we measure the entropy of the order of verbal dependents conditioned on the local subtree structure, then we will erroneously conclude that German is highly variable, since the order is either V2 or V-final and there is nothing in the local subtree to predict which one is appropriate. However, if we analyze complementizers as the dependent of their verb (as in the Universal Dependencies style, (3c)), then the conditional entropy of the verb position given local subtree structure is small. This is because the position of the verb is fully predicted by the presence in the local subtree of a *mark* relation whose dependent is *dass*, *weil*, etc.

(3c)



[1] At a high level, the bootstrap algorithm works by measuring entropy in the whole sample and in subsamples and uses these estimates to attempt to correct bias in the whole sample. We refer the reader to DeDeo et al. (2013) for details.



We deal with this issue by preferring annotation styles under which the determinants of the order of a local subtree are present in that subtree. This often means using the content-head dependency style, as in this example.

The second issue with looking only at local subtrees is that we miss certain word order variability associated with nonprojectivity, such as scrambling. Due to space constraints, we do not address this issue here.

When we condition on the local subtree structure and find the conditional entropy of word orders, we call this measure **Relation Order Entropy**, since we are getting the order with which relation types are expressed in a local subtree.

3.3 Dependency Direction

Another option for dealing with data sparsity is to get conditional entropy measures over even less dependency structure. In particular we consider the case of entropy measures conditioned only on a dependent, its head, and the relation type to its head, where the dependent measure is simply whether the head is to the left or right of the dependent. This measure potentially suffers much less from data sparsity issues, since the set of possible heads and dependents in a corpus is much smaller than the set of possible local subtrees. But in restricting our attention only to head direction, we miss the ability to measure any word order freedom among sister dependents. This measure also has the disadvantage that it can miss the kind of conditioning information present in local subtrees, as described in Section 3.2.

When we condition only on simple dependencies, we call this measure **Head Direction Entropy**.

3.4 Conditioning Variables

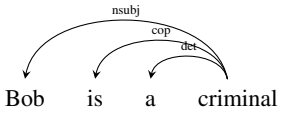
So far we have discussed our decision to use conditional entropy measures over local subtrees or single dependencies. In this setting, the conditioning variable is the unordered local subtree or dependency, and the dependent variable is the linear order of words. We now turn to the question of what information should be contained in the conditioning variable: whether it should be the full unordered tree, or just the structure of the tree, or the structure of the tree plus part-of-speech (POS) tags and relation types, etc.

In Section 3.1 we argued that we should not condition on the wordforms or lemmas due to sparsity issues. The remaining kinds of information available in corpora are the tree topology, POS tags, and relation types. Many corpora also include annotation for morphological features, but this is not reliably present.

Without conditioning on relation types, our entropy measures become much less linguistically useful. For example, if we did not condition on dependency relation types, it would be impossible to identify verbal subjects and objects or to quantify how informative word order is about these relations crosslinguistically. So we always include dependency relation type in conditioning variables.

The remaining questions are whether to include the POS tags of heads and of each dependent. Some annotation decisions in the Universal Dependencies and Stanford Dependencies argue for including POS information of heads. For example, the Universal Dependencies annotation for copular sentences has the predicate noun as the head, with the subject noun as a dependent of type *nsubj*, as in example (4):

(4)



This has the effect that the linguistic meaning of the *nsubj* relation encodes one syntactic relation when its head is a verb, and another syntactic relation when its head is a noun. So we should include POS information about heads when possible.

There are also linguistic reasons for including the POS of dependents in the conditioning variable. Word order often depends on part of speech; for example, in Romance languages, the standard order in the main clause is Subject-Verb-Object if the object is a noun but Subject-Object-Verb if the object is a pronoun. Not including POS tags in the conditioning variable would lead to misleadingly high word order freedom numbers for these clauses in these languages.

Therefore, when possible, our conditioning variables include the POS tags of heads and de-

pendents in addition to dependency relation types.

3.5 Annotation style and crosslinguistic comparability

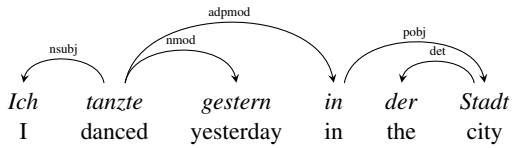
We have discussed issues involving entropy estimation and the choice of conditioning and dependent variables. Here we discuss another dimension of choices: what dependency annotation scheme to use.

Since the informativity of dependency trees about syntax and semantics affects our word order freedom measures, it is important to ensure that dependency trees across different corpora convey the same information. Certain annotation styles might allow unordered local subtrees to convey more information in one language than in another. To ensure comparability, we should use those annotation styles which are most consistent across languages regarding how much information they give about words in local subtrees, even if this means choosing annotation schemes which are less informative overall. We give examples below.

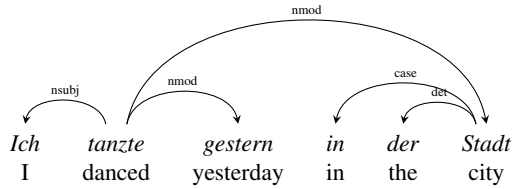
In many cases, dependency annotation schemes where function words are heads provide more information about syntactic and semantic relations, so such annotation schemes lead to lower estimates of word order freedom. For example, consider the ordering of German verbal adjuncts. The usual order is time adjuncts followed by place adjuncts. Time is often expressed by a bare noun such as *gestern* "yesterday", while place is often expressed with an adpositional phrase.

We will consider how our measures will behave for these constructions given function-word-head dependencies, and given content-head dependencies. Given function-word-head dependencies as in (5a), these two adjuncts will appear with relations *nmod* and *adpmod* in the local subtree rooted by the verb *tanzte*; their order will be highly predictable given these relation types inasmuch as time adjuncts are usually expressed as bare nouns and place adjuncts are usually expressed as adpositional phrases. On the other hand, given content-head dependencies as in (5b), the adjuncts will appear in the local subtree as *nmod* and *nmod*, and their order will appear free.

(5a)



(5b)



However, function-word-head dependencies do not provide the same amount of information from language to language, because languages differ in how often they use adpositions as opposed to case marking. In the German example, function-word-head dependencies allowed us to distinguish time adjuncts from place adjuncts because place adjuncts usually appear as adpositional phrases while time adjuncts often appear as noun phrases. But in a language which uses case-marked noun phrases for such adjuncts, such as Finnish, the function-word-head dependencies would not provide this information. Therefore, even if (say) Finnish and German had the same degree of freedom in their ordering of place adjuncts and time adjuncts, we would estimate more word order freedom in Finnish and less in German. However, using content-head dependencies, we get the same amount of information in both languages. Therefore, we prefer content-head dependencies for our measures.

Following similar reasoning, we decide to use only the universal POS tags and relation types in our corpora, and not finer-grained language-specific tags.

Using content-head dependencies while conditioning only on local subtrees overestimates word order freedom compared to function-word-head dependencies. At first glance, the content-head dependency annotation seems inappropriate for a typological study, because it clashes with standard linguistic analyses where function words such as adpositions and complementizers (and, in some analyses, even determiners (Abney, 1987)) are heads, rather than dependents. However, content-head dependencies provide more consistent measures across languages. Therefore we present results from our measures applied to content-head dependencies.

3.6 Summary of Parameters of Entropy Measures

We have discussed a number of parameters which go into the construction of a conditional entropy

measure of word order freedom. They are:

1. Annotation style: function words as heads or content words as heads.
2. Whether we measure entropy of linearizations of local subtrees (*Relation Order Entropy*) or of simple dependencies (*Head Direction Entropy*).
3. What information we include in the conditioning variable: relation types, head and dependent POS, head and dependent wordforms, etc.
4. Whether to measure entropy over all dependents, or only over some subset of interest, such as subjects or objects.

The decisions for these parameters are dictated by balancing data sparsity and linguistic interpretability. We have argued that we should use content-head dependencies, and never include wordforms or lemmas in the conditioning variables. Furthermore, we have argued that it is generally better to include part-of-speech information in the conditioning variable, but that this may have to be relaxed to cope with data sparsity. The decisions about whether to condition on local subtrees or on simple dependencies, and whether to restrict attention to a particular subset of dependencies, depends on the particular question of interest.

3.7 Entropy Measures as Upper Bounds on Word Order Freedom

We initially defined an ideal measure, the entropy of word orders given full unordered dependency trees. We argued that we would have to back away from this measure by looking only at the conditional entropy of orders of local subtrees, and furthermore that we should only condition on the parts of speech and relation types in the local subtree. Here we argue that these steps away from the ideal measure mean that the resulting measures can only be interpreted as upper bounds on word order freedom.

With each step away from the ideal measure, we also move the *interpretation* of the measures away from the idealized notion of word order freedom. With each kind of information we remove from the independent variable, we allow instances where the word order of a phrase might in fact be fully deterministic given that missing information, but where we will erroneously measure high word order freedom. For example, in German, the order of verbal adjuncts is usually time before place. However, in a dependency treebank, these relations are all *nmod*. By considering only the ordering of dependents with respect to their relation types and parts of speech, we miss the extent to which these dependents *do* have a deterministic order determined by their semantics. Thus, we tend to overestimate true word order freedom.

On the other hand, the conditional entropy approach do not in principle *under*estimate word order freedom as we have defined it. The conditioning information present in a dependency tree represents only semantic and syntactic relations, and we are explicitly interested in word order variability beyond what can be explained by these factors. Therefore, our word order freedom measures constitute upper bounds on the true word order freedom in a language.

Underestimation can arise due to data sparsity issues and bias issues in entropy estimators. For this reason, it is important to ensure that our measures are stable with respect to sample size, lest our upper bound become a lower bound on an upper bound.

The tightness of the upper bound on word order freedom depends on the informativity of the relation types and parts of speech included in a measure. For example, if we use a system of relation types which subdivides *nmod* relations into categories like *nmod:tmod* for time phrases, then we would not overestimate the word order freedom of German verbal adjuncts. As another example, to achieve a tighter bound for a limited aspect of word order freedom at the cost of empirical coverage, we might restrict ourselves to relation types such as *nsubj* and *dobj*, which are highly informative about their meanings.

4 Applying the Measures

Here we give the results of applying some of the measures discussed in Section 3 to dependency corpora. We use the dependency corpora of the HamleDT 2.0 (Zeman et al., 2012; Rosa et al., 2014) and Universal Dependencies 1.0 (Nivre et al., 2015). All punctuation and dependencies with relation type *punct* are removed. We only examine sentences with a single root. Annotation was normalized to content-head format when necessary. Combined this gives us dependency corpora of 34 languages in a fairly standardized format.

In order to evaluate the stability of our measures with respect to sample size, we measure all en-

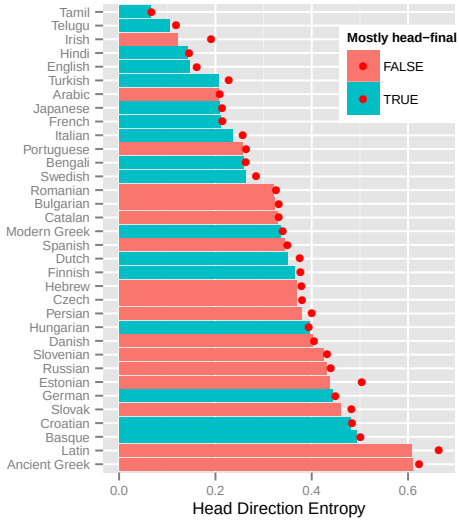


Figure 1: Head direction entropy in 34 languages. The bar represents the average magnitude of head direction entropy estimated from subcorpora of 1000 sentences; the red dot represents head direction entropy estimated from the whole corpus.

tropies using the bootstrap estimator of DeDeo et al. (2013). We report the mean results from applying our measures to subcorpora of 1000 sentences for each corpus. We also report results from applying measures to the full corpus, so that the difference between the full corpus and the subcorpora can be compared, and the effect of data sparsity evaluated.

4.1 Head Direction Entropy

Head direction entropy, defined and motivated in Section 3.3, is the conditional entropy of whether a head is to the right or left of a dependent, conditioned on relation type and part of speech of head and dependent. This measure can reflect either consistency in head direction conditioned on relation type, or consistency in head direction *overall*. Results from this measure are shown in Figure 1. As can be seen, the measure gives similar results when applied to subcorpora as when applied to full corpora, indicating that this is measure is not unduly affected by differences in sample size.

We find considerable variability in word order freedom with respect to head direction. In languages such as Korean, Telugu, Irish, and English, we find that head direction is nearly deterministic. On the other hand, in Slavic languages and in Latin and Ancient Greek we find great variability. The fact that entropy measures on subcorpora of 1000 sentences do not diverge greatly from entropy measures on full corpora indicates that this measure is stable with respect to sample size.

We find a potential relationship between predominant head direction and word order freedom in head direction. Figure 1 is coded according to whether languages have more than 50% head-final dependencies or not. The results suggest that languages which have highly predictable head direction might tend to be mostly head-final languages.

The results here also have bearing on appropriate generative models for grammar induction. Common generative models, such as DMV, use separate multinomial models for left and right dependents of a head. Our results suggest that for some languages there should be some sharing between these distributions.

4.2 Relation Order Entropy

Relation order entropy (Section 3.2) is the conditional entropy of the order of words in a local subtree, conditioned on the tree structure, relation types, and parts of speech. Figure 2 shows relation order entropy for our corpora. As can be seen, this measure is highly sensitive to sample size: for corpora with a medium sample size, such as English (16535 sentences), there is a moderate difference between the results from subcorpora and the results from the full corpus. For other languages with comparable size, such as Spanish (15906 sentences), there is a larger difference. In the case of languages with small corpora such as Bengali (1114 sentences), their true relation order entropy is almost certainly higher than measured.

While relation order entropy is the most easily interpretable and general measure of word order freedom, it does not seem to be workable given current corpora and methods. In further experiments, we found that removing POS tags from the conditioning variable does not reduce the instability of this measure.

4.3 Relation Order Entropy of Subjects and Objects

We can alleviate the data sparsity issues of relation order entropy by restricting our attention to a few relations of interest. For example, the position of subject and object in the main clause has long been of interest to typologists (Greenberg, 1963), (cf. (Dryer, 1992)). In Figure 3 we present relation order entropy of subject and object for local subtrees containing relations of type *nsubj* and *dobj* (*obj* in

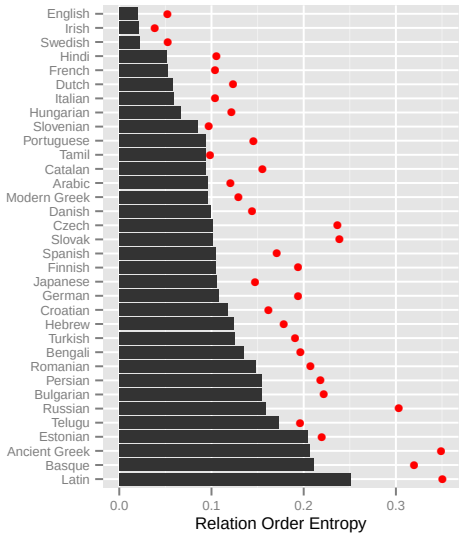


Figure 2: Relation order entropy in 34 languages. The bar represents the average magnitude of relation order entropy estimated from subcorpora of 1000 sentences; the red dot represents relation order entropy estimated from the whole corpus.

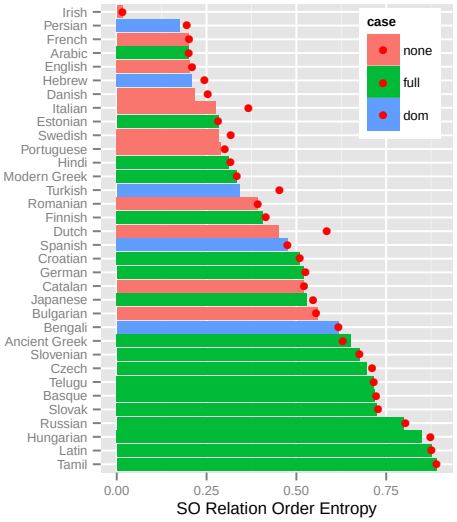


Figure 3: Relation order entropy for subject and object in 34 languages. Language names are annotated with corpus size in number of sentences. Bars are colored depending on the nominative-accusative case marking system type for each language. "Full" means fully present case marking in at least one paradigm. "dom" means Differential Object Marking.

the case of HamleDT corpora), conditioned on the parts of speech for these dependents.

The languages Figure 3 are colored according to their nominative-accusative[2] case marking on nouns. We consider a language to have full case marking if it makes a consistent morphological distinction between subject and object in at least one paradigm. If the distinction is only present conditional on animacy or definiteness, we mark the language as DOM for Differential Object Marking (Aissen, 2003).

The figure reveals a relationship between morphology and this particular aspect of word order freedom. Languages with relation order entropy above .625 all have relevant case marking, so it seems word order freedom in this domain implies the presence of case marking. However, case marking does not imply rigid word order; several languages in the sample have rigid word order while still having case marking. Our result is a quantitative sharpening of the pattern claimed in Kiparsky (1997).

Interestingly, many of the exceptional languages—those with case marking and rigid word order—are languages with verb-final or verb-initial orders. In our sample, Persian, Hindi, and Turkish are case-marking verb-final languages where we measure low levels of freedom in the order of subject and object. Modern Standard Arabic is (partly) verb-initial and case-marking (although case marking is rarely pronounced or explicitly written in modern Arabic). This finding is in line with recent work (Gibson et al., 2013; Futrell et al., 2015) which has suggested that verb-final and verb-initial orders without case marking do not allow robust communication in a noisy channel, and so should be dispreferred.

5 Conclusion

We have presented a set of interrelated methodological and linguistic issues that arise as part of quantifying word order freedom in dependency corpora. We have shown that conditional entropy measures can be used to get reliable estimates of variability in head direction and in ordering relations for certain restricted relation types. We have argued that such measures constitute upper bounds on word order freedom. Further, we have demonstrated a simple relationship between morphological case marking and word order freedom in the domain of subjects and objects, providing to our

[2]Or ergative-absolutive in the case of Basque and the Hindi past tense.

knowledge the first large-scale quantitative validation of the old intuition that languages with free word order must have case marking.

Acknowledgments

We thank Leon Bergen, Paola Merlo, and the audience at AMLaP 2014 for helpful discussion. K.M. was supported by the Department of Defense through the National Defense Science & Engineering Graduate Fellowship program.

References

Steven Paul Abney. 1987. *The English noun phrase in its sentential aspect*. Ph.D. thesis, MIT.

Olga Abramov and Alexander Mehler. 2011. Automatic language classification by means of syntactic dependency networks. *Journal of Quantitative Linguistics*, 18(4):291–336.

Judith Aissen. 2003. Differential object marking: Iconicity vs. economy. *Natural Language & Linguistic Theory*, 21(3):435–483.

Franklin Chang. 2009. Learning to order words: A connectionist model of Heavy NP Shift and accessibility effects in Japanese and English. *Journal of Memory and Language*, 61:374–397.

Greville G Corbett. 2006. *Agreement*. Cambridge University Press.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal stanford dependencies: A cross-linguistic typology. In *Proceedings LREC'14*.

Simon DeDeo, Robert X. D. Hawkins, Sara Klingenstein, and Tim Hitchcock. 2013. Bootstrap methods for the empirical study of decision-making and information flows in social systems. *Entropy*, 15(6):2246–2276.

Matthew S Dryer. 1992. The greenbergian word order correlations. *Language*, 68(1):81–138.

Victor S Ferreira and Hiromi Yoshita. 2003. Given-new ordering effects on the production of scrambled sentences in Japanese. *Journal of Psycholinguistic Research*, 32(6):669–692.

Richard Futrell, 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*, 136:215–221.

Edward Gibson, 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(7):1079–88.

Joseph Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph Greenberg, editor, *Universals of Language*, pages 73–113. MIT Press.

Paul Kiparsky. 1997. The rise of positional licensing. In Ans von Kemenade and Nigel Vincent, editors, *Parameters of morphosyntactic change*, pages 460–494. Cambridge University Press.

Dan Klein and Christopher D Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the ACL*, page 478.

Marco Kuhlmann. 2013. Mildly non-projective dependency grammar. *Computational Linguistics*, 39(2):355–387.

Haitao Liu and Wenwen Li. 2010. Language clusters based on linguistic complex networks. *Chinese Science Bulletin*, 55(30):3458–3465.

Haitao Liu. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120(6):1567–1578.

Thomas McFadden. 2003. On morphological case and word-order freedom. In *Proceedings of the Berkeley Linguistics Society*.

George Miller. 1955. Note on the bias of information estimates. In *Information Theory in Psychology: Problems and Methods*, pages 95–100.

Johanna Nichols. 1986. Head-marking and dependent-marking grammar. *Language*, 62(1):56–119.

Joakim Nivre et al. 2015. *Universal Dependencies 1.0*. Universal Dependencies Consortium.

Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In *Computational Linguistics and Intelligent Text Processing*, pages 3–16. Springer.

Rudolf Rosa, Jan Maek, David Mareek, Martin Popel, Daniel Zeman, and Zdenk abokrtsk. 2014. HamleDT 2.0: Thirty dependency treebanks Stanfordized. In *Proceedings LREC'14*.

Edward Sapir. 1921. *Language, an introduction to the study of speech*. Harcourt, Brace and Co., New York.

Daniel Zeman, David Marecek, Martin Popel, Loganathan Ramasamy, Jan Stepánek, Zdenek Zabokrtskỳ, and Jan Hajic. 2012. HamleDT: To parse or not to parse? In *Proceedings LREC'12*.

George Kingsley Zipf. 1949. *Human behavior and the principle of least effort*. Addison-Wesley Press, Oxford, UK.