Syntactic dependencies correspond to word pairs with high mutual information

Richard Futrell[1], **Peng Qian**[2], **Edward Gibson**[2], **Evelina Fedorenko**[2], **and Idan Asher Blank**[3]
[1] Department of Language Science, University of California, Irvine
[2] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
[3] Department of Psychology, University of California, Los Angeles
rfutrell@uci.edu, {pqian, egibson, evelina9}@mit.edu, iblank@psych.ucla.edu

Abstract

How is syntactic dependency structure reflected in the statistical distribution of words in corpora? Here we give empirical evidence and theoretical arguments for what we call the Head–Dependent Mutual Information (HDMI) Hypothesis: that syntactic heads and their dependents correspond to word pairs with especially high mutual information, an information-theoretic measure of strength of association. In support of this idea, we estimate mutual information between word pairs in dependencies based on an automatically-parsed corpus of 320 million tokens of English web text, finding that the mutual information between words in dependencies is robustly higher than a controlled baseline consisting of non-dependent word pairs. Next, we give a formal argument which derives the HDMI Hypothesis from a probabilistic interpretation of the postulates of dependency grammar. Our study also provides some useful empirical results about mutual information in corpora: we find that maximum-likelihood estimates of mutual information between raw wordforms are biased even at our large sample size, and we find that there is a general decay of mutual information between part-of-speech tags with distance.

1 Introduction

The field of quantitative syntax requires a way to link the discrete formal structures typically studied in syntax, such as dependency trees, with the probabilistic distributions over wordforms observable in corpora.

Formal syntactic structures are usually taken to define the categorical well-formedness of sentences (Chomsky, 1957), or the latent structures required to derive an interpretation (Heim and Kratzer, 1998). It remains unclear what relationship should obtain between these structures and statistical co-occurrence patterns over linguistic units as one might observe in a corpus. Early work in linguistics tried to use these co-occurrence patterns as the basis on which to define formal syntactic structures, formulating 'discovery procedures' which would enable co-occurrence statistics to be summarized mechanistically using formal syntactic structures (Harris, 1954), but modern generative theories of syntax have eschewed any connection between statistical and syntactic structure (Adger, 2018), and to date it remains unclear whether corpus statistics contain enough information to fully reconstruct syntactic structures as identified by linguists. NLP researchers working on grammar induction and unsupervised parsing have achieved substantial gains in recovering dependency trees on the basis of corpus statistics, but overall accuracy remains modest (Klein and Manning, 2004; Spitkovsky et al., 2012; Le and Zuidema, 2015; Pate and Johnson, 2016; Jiang et al., 2016).

Here we propose a high-level linking hypothesis between dependency structures and co-ocurrence statistics: syntactic dependencies correspond to word pairs with high **mutual information (MI)**, an information-theoretic measure of the strength of covariance between two random variables (Cover and Thomas, 2006). We call this claim the **Head–Dependent Mutual Information (HDMI) Hypothesis**. In doing so we formalize and justify an intuition that has underlain much of the work on grammar induction for over 20 years (de Paiva Alves, 1996; Yuret, 1998; Klein and Manning, 2004). The basic intuition is that MI is a generic measure of strength of covariance, and heads and dependents are those word pairs whose covariance is most strongly constrained by grammatical rules.

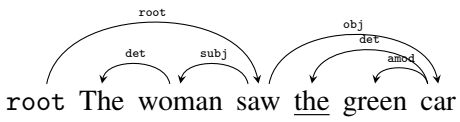

Figure 1: Example dependency tree.

The paper is structured as follows. Sections 2 discusses the methods and dataset we used to measure MI and evaluate the HDMI Hypothesis in corpora; we believe this to be the largest-scale attempt to date to estimate MI between wordforms in natural language text. In Section 3, we present the results of the study, showing that dependencies do identify word pairs with especially high MI as measured in various ways. We also find that mutual information between part-of-speech tags decreases with distance, but we do not observe a similar decay pattern for mutual information between words represented as distributional clusters. Next, in Section 4, we elaborate on the theoretical justification for the HDMI Hypothesis, providing a formal derivation of the hypothesis from an information-theoretic interpretation of the basic postulates of dependency grammar. Section 5 concludes.

2 Measuring Head–Dependent MI in Dependency Corpora

We evaluate the HDMI Hypothesis in a large automatically-parsed corpus of English. To do so, we calculate mutual information between heads and dependents in the corpus. For example, Figure 1 shows an example of a dependency tree. The tree has five **dependency pairs**: ordered pairs of words where the first element is a head and the second is its dependent. The dependency pairs based on this tree are <saw, woman>, <woman, the>, <saw, car>, <car, green>, and <car, the> (excluding the `root` dependency). We calculate Head–Dependent Mutual Information (HDMI) between heads h and dependents d in these pairs:

$$\text{HDMI} = \mathbb{E}\left[\log \frac{p(h,d)}{p(h)p(d)}\right].$$

As a baseline, we also compare HDMI against the mutual information of pairs of words that are not in a direct dependency relationship. See Section 2.2 for details on how these non-dependency word pairs are selected.

For evidence for the HDMI Hypothesis from other languages and hand-parsed corpora, see Futrell and Levy (2017). To our knowledge, the current work is the largest-scale attempt to date to estimate mutual information between words in natural language text and to demonstrate the relationship between dependency and mutual information in a controlled way. The code for our analysis can be found online at `http://github.com/pqian11/mi-hdmi`.

2.1 Estimating Mutual Information

We estimate mutual information using maximum likelihood estimation applied to joint count data over wordforms. The mutual information between wordforms is the true mutual information of interest for our hypothesis, but it is not clear that we can achieve accurate estimates of this quantity due to data sparsity. Therefore we also calculate mutual information between part-of-speech (POS) tags and between distributional clusters, described in more detail below. We include all dependencies except the `root` dependency and those involving wordforms that are not among the top 60,000 most frequent wordforms in the whole corpus.

These experiments also provide data on the convergence of mutual information estimates for wordforms. It is notoriously challenging to estimate information-theoretic quantities such as entropy and mutual information from count data (Miller, 1955; Paninski, 2003; Archer et al., 2013), especially for distributions with long tails, such as wordforms of natural language. Bentz et al. (2017) show that word-level entropy estimates, calculated using maximum likelihood estimation (MLE), converge with around 10^5 tokens of text. But estimating mutual information is more challenging because it requires estimating

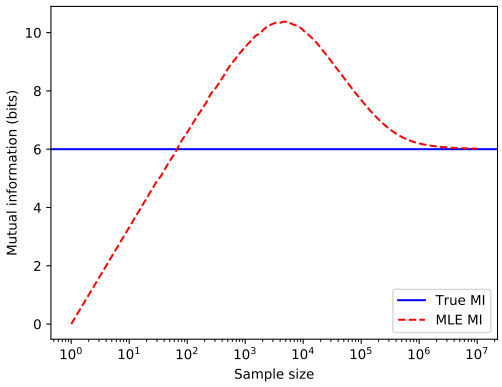

Figure 2: Illustration of bias in MLE estimates of mutual information. The example distribution here is a joint distribution over pairs of bitstrings of length 12, where each pair shares 6 bits, so true mutual information is equal to 6 bits by construction. Empirical MLE estimates of mutual information are shown for various sample sizes. The mutual information estimate initially underestimates the true value, then overestimates it before eventually approaching the true value at around 10^7 samples.

a joint distribution over pairs of words, not just a distribution over single words. It is therefore unknown at what sample size mutual information estimates would converge. Furthermore, while MLE estimates of entropy have a general downward bias, the bias of mutual information is not necessarily downward or upward, as shown in Figure 2. Therefore the MI estimation problem is harder than the entropy estimation problem, because we might not know for some sample size whether we are in an underestimation phase or an overestimation phase.

2.2 Matched non-dependency baseline

We compare the MI of words in dependencies against the MI of words in a **matched non-dependency baseline**. These are word pairs which are not in a direct dependency relationship, and which are matched with the dependency pairs in terms of **displacement**: the linear distance from the head to the dependent and the direction of the dependent with respect to the head, calculated as the linear index of the dependent minus the linear index of the head. For example, given the tree in Figure 1, we might take the non-dependency word pairs <green, the> (displacement -1), <the, saw> (displacement -1), <woman, green> (displacement 3), and <green, saw> (displacement -2). We collect the same number of non-dependency word pairs as dependency word pairs from the corpus. We predict higher MI among the dependency word pairs than among these baseline word pairs.

2.3 Permuted baseline

In order to quantify the magnitude of estimation bias affecting our results, we also compute mutual information for a baseline case where we shuffle the mapping between observed heads and dependents for the entire corpus. In this **permuted baseline**, heads and dependents have analytically zero mutual information: the shuffling process destroys all covariance between heads and dependents within sentences. If our estimation procedures yield any mutual information at all in this case, it can only be due to data sparsity. Therefore the shuffled baseline provides a measure of the strength of the bias affecting our estimates.

2.4 Statistical tests

We wish to statistically compare the MI of dependency word pairs against the MI of the matched non-dependency baseline. To do so, we need some measure of the variance in our MI estimates. Therefore we split our data into 16 equally-sized subsets and calculate MI separately within each subset, and use the standard error of the resulting 16 data points to calculate 95% confidence intervals for each MI estimate. In all figures below, except where otherwise noted, each displayed MI values is the mean of MI values

obtained from the 16 subsets. The confidence intervals are too small to be seen. To compare two mean MI estimates statistically, we used two-tailed paired t-tests and report p-values following a False Discovery Rate correction for multiple comparisons (Benjamini and Yekutieli, 2001).

2.5 Dataset

We use the Common Crawl corpus (Buck et al., 2014) of English web text. We filtered the corpus to contain mostly meaningful linguistic utterances and to remove irrelevant web boilerplate text.[1] We parsed and POS-tagged 10% of the filtered corpus using SyntaxNet (Andor et al., 2016). The final dataset used in this paper consists of a total of 320 million tokens of parsed text. SyntaxNet produces function-word-headed dependencies, rather than content-head dependencies, so our results reflect syntactic dependencies rather than semantic dependencies.

2.5.1 POS tags

For MI between POS tags, we use the Penn Treebank POS tags output by SyntaxNet.

POS tags can be interpreted *roughly* as a lower bound on the true MI between full wordforms, because POS tags are mostly a function of individual wordforms. This interpretation is rough because POS tags are to some extent context-dependent. In any case, they can be interpreted as representing the syntactic information present in a word token.

2.5.2 Distributional clusters

Our distributional clusters are derived by spectral clustering from the 300-dimensional GloVe word embedding space trained on 42 billion tokens of the uncased English Common Crawl corpus (Pennington et al., 2014). To generate distributional clusters, we first select the most frequent 60,000 words from a chunk of the whole corpus. After filtering out words that do not have a pretrained embedding in GloVe, we compute the similarity matrix for the remaining 59,998 words and run a spectral clustering algorithm based on the similarity matrix (Pedregosa et al., 2011).

We derive 300 clusters by this method. We found empirically that going above around 300 clusters resulted in many singleton clusters.

We calculate MI between distributional clusters by replacing each word with the index of its cluster and then computing MI by MLE between co-occurrence counts of cluster indices. Because the distributional cluster for a word is a function only of its wordform, the MI between distributional clusters is a true lower bound on MI between full wordforms.

3 Results

3.1 Convergence of MI estimates

Figure 3 shows the convergence of MI estimates for wordforms with increasing sample size, for dependency pairs, matched non-dependency pairs, and the permuted baseline. We see that MI is systematically overestimated at small sample sizes, and that the estimates decrease with increasing sample size. However, even with sample sizes on the order of 10^7 to 10^8 tokens, the estimate does not appear to have converged to a stable value. Furthermore, we see that the permuted baseline, which should ultimately converge to an estimate of zero, still yields a substantial positive MI estimate (0.47 bits) given the full corpus. Overall, we conclude that it is not possible to get an unbiased and stable estimate of MI between wordforms with 10^8 or fewer tokens of text using maximum likelihood estimation. More accurate estimates could come from larger data or from more sophisticated methods of estimating MI.

Now we turn to the convergence of MI estimates based on POS tags and distributional clusters, as shown in Figure 4. Estimates based on POS tags appear to have already converged at 10^5 tokens, and estimates based on distributional clusters appear to converge around 10^7 tokens. Furthermore, the permuted

[1] We filtered out all lines that did not begin with a capital letter and end with punctuation, and all lines containing "copyright", "download", "error", or days of the week or names of months: these lines were overwhelmingly boilerplate text. The filtration process removed about 90% of lines.

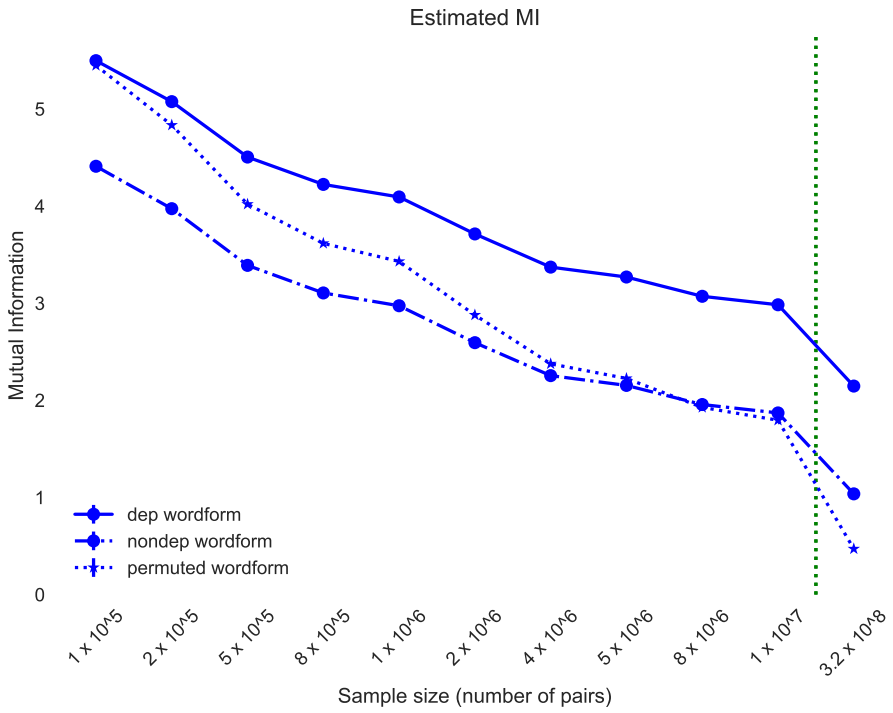

Figure 3: Maximum likelihood-estimated MI by sample size, for wordforms in dependencies (dep), matched non-dependencies (nondep), and the permuted baseline (permuted). The points to the left of the green line are average MI values from 16 subsets of the data. The points to the right of the green line are single point estimates computed from the full corpus.

baseline is near zero for the estimates based on POS tags, and eventually drops to near zero for distributional clusters, an encouraging result that indicates that we have sufficient data to overcome estimation bias due to data sparsity.

3.2 HDMI Hypothesis

Figures 3 and 4 already show that MI in dependencies is higher than in non-dependencies, supporting the HDMI Hypothesis, for POS tags, distributional clusters, and raw wordforms (although the estimation bias for the latter makes the interpretation difficult). For raw wordforms, the difference between dependency MI and baseline MI is significant in all sample sizes at $p < 10^{-16}$.

3.3 Decay with distance

The relationship between mutual information and distance is of theoretical interest beyond the HDMI Hypothesis. Li (1989) and Lin and Tegmark (2017) have reported that mutual information between orthographic letters in natural language text falls off as a power law with distance, but the relationship between mutual information and distance at the level of words has not yet been explored in large corpora. Because of the estimation difficulties observed in Section 3.1, we do not analyze MI between raw wordforms here, but rather only between POS tags and distributional clusters.

We estimate mutual information between POS tags and distributional at different distances. We hold sample size constant for all distances, meaning that we have around 5×10^6 dependency pairs available to estimate MI at each distance. Figure 5 shows the results. We see a clear fall-off of mutual information with distance for POS tags, for both dependencies and non-dependencies. However, we see no fall-off for distributional clusters, indicating that it may be primarily syntactic information that drives high-MI words to be close to each other.

We can also see from Figure 5 that the HDMI Hypothesis holds for both POS tags and distributional clusters in all distances shown (with significance at $p < 10^{-18}$ at all distances shown).

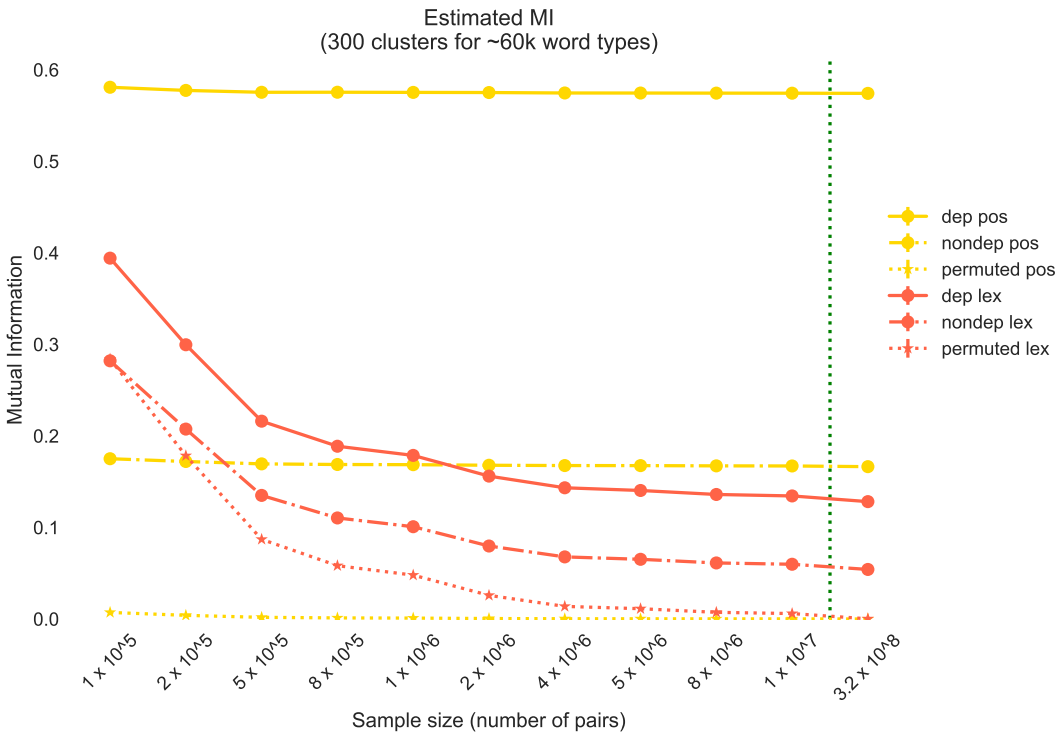

Figure 4: Maximum likelihood-estimated MI by sample size, for wordforms in dependencies (dep), matched non-dependencies (nondep), and the permuted baseline (permuted), based on POS tags (yellow) and distributional clusters (red). Green line as in Figure 3.

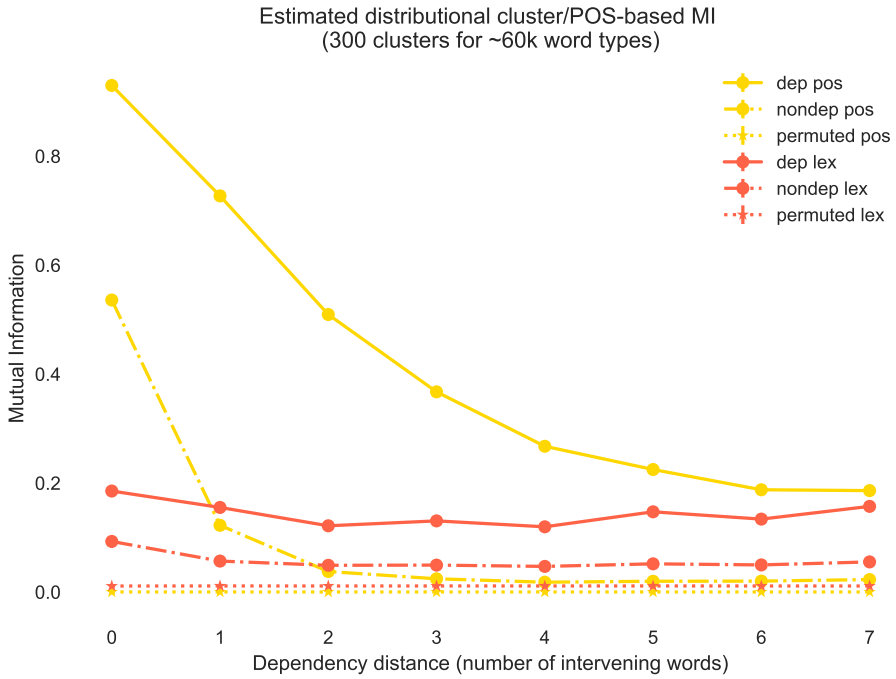

Figure 5: Estimated mutual information between words by number of intervening words (dependency distance), as estimated using POS tags (yellow) and distributional clusters (red). MI for each distance is estimated from $\sim$ 5 million dependency pairs, averaged over 16 subsets of the corpus.

One caveat is in order regarding the interpretation of Figure 5: these results are based on an automatically-parsed corpus, and the more distant dependencies may be less accurately identified by the parser. Long dependencies are known to cause difficulties for shift–reduce parsers such as SyntaxNet (Gulordava and Merlo, 2015). Therefore in our dataset, it may be the case that the longer dependencies are noisier, and their MI will thus regress to the MI of non-dependency word pairs. However, we note that Futrell and Levy (2017) found a similar decay of POS tag MI with distance even in hand-parsed corpora.

4 Theoretical justification

Having established that the HDMI Hypothesis holds empirically in large corpora, we now turn to a theoretical justification for this hypothesis. We propose to view dependency grammar as a method for approximating arbitrary probability distributions over strings. Taking this view, we show that choosing dependency trees to minimize approximation error is equivalent to choosing dependency trees to maximize the head–dependent MI. Therefore the most accurate dependency trees in a linguistic sense will be those with maximal head–dependent MI.

We take the basic postulate of dependency grammar to be that the syntactic well-formedness of a sentence can be fully or mostly characterized in terms of the pairs of head and dependent words in the sentence as identified by some dependency tree (Hudson, 1984, 2010). That is, restrictions on covariance between words in sentences can be stated entirely in terms of the head–dependent pairs forming dependency trees. Given a dependency tree such as the one in Figure 1, all that you would need to know to specify the conditions on what word can go in the underlined position is the identity of the head word—*car*, a noun, licensing a determiner as a dependent. There may also be dependency type labels which are relevant. In the strongest possible formulation of dependency grammar—undoubtedly too strong—the head provides *all* the information you need to specify the possible dependents. More realistically, we can say that the identities of other words in the dependency tree, which are only distantly connected to the underlined word in terms of the dependency structure, play relatively minor roles in the determination of the underlined word.

While dependency grammar was developed to specify categorical well-formedness conditions, we can make a probabilistic generalization and say that the *probability* of a sentence is fully or mostly characterizable in terms of the head and dependent pairs. This assumption is closely related to generative models from the grammar induction literature called **head-outward generative models** (Eisner, 1996; Klein and Manning, 2004), in which the probability of a sentence can be factorized in terms of a dependency tree. Representing a sentence as a sequence of n words $\mathbf{w}_{i=1}^{n}$, and representing the dependency tree for the sentence as a sequence of n heads $\mathbf{t}_{i=1}^{n}$ where t_i gives the head of the ith word w_i, we can factorize the probability of the sentence $\mathbf{w}$ as:

$$p_{\mathbf{t}}(\mathbf{w}_{i=1}^{n}) = \prod_{i=1}^{n} p_{\mathbf{t}}(w_i|t_i).^{2} \tag{1}$$

We propose to view dependency grammar in this sense as a generic method for approximation to arbitrary distributions over sequences, closely related to Chow-Liu trees (Chow and Liu, 1968), which are a general scheme for approximating any joint distribution in terms of only pairwise dependencies.

Any distribution over sequences of symbols could be approximated by Eq. 1 for some set of dependency trees specified by $\mathbf{t}$, to varying degrees of accuracy. Eq. 1 fundamentally expresses an assumption that all the relevant information in the context about the symbol w_i is concentrated in exactly one other symbol t_i—corresponding to the dependency grammar postulate described above.

The independence assumptions of Eq. 1 are obviously too strong for natural language, but surprisingly they provide a reasonable approximation (Eisner, 1996, 1997). It is an interesting scientific question why natural language has the property that it can be well-approximated by such a dependency grammar.

[2]Head-outward generative models differ from our Eq. 1 in that they also put a prior distribution over tree structures $\mathbf{t}$. In contrast, we are only interested in the probability of a string $\mathbf{w}$ given a tree $\mathbf{t}$. One consequence of our formulation is that the *halting probabilities* that appear in Eisner (1996) do not appear in our equations. Since we are not considering prior probabilities on tree structures, our approach in this section is similar to fitting a head-outward generative model by maximum likelihood estimation.

We propose that when linguists are developing dependency grammars and assigning dependency trees to sentences, they are implicitly finding trees $\mathbf{t}$ to make the approximation in Eq. 1 as accurate as possible. That is, for each word, they are choosing the heads that best explain the distribution of each word in the sentence. More formally, they are solving the problem of minimizing the divergence between the dependency approximation in Eq. 1 and the true distribution over sequences of symbols (sentences), which we call p_L. The true distribution over sequences p_L can be written generically as:

$$p_L(\mathbf{w}_{i=1}^n) = \prod_{i=1}^{n} p_L(w_i|\mathbf{w}_{<i}), \tag{2}$$

where $\mathbf{w}_{<i}$ represents the sequence of symbols up to the ith (non-inclusive).

We now show that minimizing the KL-divergence between the true distribution over sequences p_L (Eq. 2) and the dependency approximation $p_\mathbf{t}$ (Eq. 1) is equivalent to choosing head–dependent pairs that maximize mutual information. This result provides a conceptual link between dependency grammar and information-theoretic statistics observable in corpora.

More formally, let p_L be a conditional probability distribution with support over symbols w_i, called **words**, and a special sentinel symbol which marks the end of a sentence. The distribution p_L generates symbols conditional on a sequence of previous words $\mathbf{w}_{<i}$, called a **context**, also generated by p_L, and starting with a special beginning-of-sentence symbol called `root`. Let $\mathbf{t}$ be a sequence of symbols, called **heads**, where t_i is equal to some word w_j for $j < i$ or to `root`, such that the pairs $\langle t_i, w_i \rangle$ define a dependency graph within each sentence.[3] We hold the distribution p_L to be a fixed target, and we are interested in finding the assignment of heads $\mathbf{t}$ that minimizes the expected per-symbol KL-divergence between the dependency approximation $p_\mathbf{t}$ of L and the true distribution p_L:

$$D_{\mathrm{KL}}(p_L(w_i|\mathbf{w}_{<i})||p_\mathbf{t}(w_i|t_i)) = \mathbb{E}\left[\log \frac{p_L(w_i|\mathbf{w}_{<i})}{p_\mathbf{t}(w_i|t_i)}\right]. \tag{3}$$

Proposition 1. *The heads* $\mathbf{t}$ *that minimize approximation error (Eq. 3) are given by:*

$$\operatorname*{argmax}_{\mathbf{t}} I[W:T],$$

where W *is the distribution over single words generated by* p_L*,* T *is the distribution over elements of* $\mathbf{t}$*, and* $I[W:T]$ *gives the mutual information of* W *and* T*, called the* ***Head–Dependent Mutual Information (HDMI):***

$$I[W:T] = \mathbb{E}\left[\log \frac{p_\mathbf{t}(w_i|t_i)}{p(w_i)}\right].$$

Proof. We begin by applying Bayes' rule to the numerator of the log probability ratio in Eq. 3:

$$\begin{aligned} D_{\mathrm{KL}}(p_L(w_i|\mathbf{w}_{<i})||p_\mathbf{t}(w_i|t_i)) &= \mathbb{E}\left[\log \frac{p_L(w_i|\mathbf{w}_{<i})}{p_\mathbf{t}(w_i|t_i)}\right] \\ &= \mathbb{E}\left[\log \frac{p(\mathbf{w}_{<i}|w_i)p(w_i)}{p(\mathbf{w}_{<i})p_\mathbf{t}(w_i|t_i)}\right]. \end{aligned} \tag{3}$$

Now we separate the result into two terms:

$$\min_{\mathbf{t}} D_{\mathrm{KL}}(p_L(w_i|\mathbf{w}_{<i})||p_\mathbf{t}(w_i|t_i)) = \min_{\mathbf{t}} \mathbb{E}\left[\log \frac{p(w_i)}{p_\mathbf{t}(w_i|t_i)}\right] + \cancel{\mathbb{E}\left[\log \frac{p(\mathbf{w}_{<i}|w_i))}{p(\mathbf{w}_{<i})}\right]}. \tag{4}$$

[3] Our construction includes an assumption that t_i for each word w_i is equal to some previous word $w_{j<i}$. This assumption may appear to entail that our dependency trees are strictly head-initial. However, the assumption is without loss of generality, because the order of the indices in Eq. 2 is arbitrary and does not have to correspond to the linear order of words: different orders simply correspond to different applications of the chain rule for probabilities and will yield the same total probability, as long as the context $\mathbf{w}_{<i}$ is encoded in such a way that the original indices are recoverable. Therefore it is always possible to reassign indices within a sentence such that the dependency graph defined by $\mathbf{t}$ appears to be strictly head-initial, while the value of Eq. 2 will remain the same. Similarly, the second term in Eq. 4 below is also invariant to the choice of indices, because it is equivalent to the average MI between contexts and words. So our result will hold for all tree structures within sentences, be they head-initial, head-final, or mixed within sentences.

The last term in Eq. 4 is the mutual information of words with their contexts under p_L. This quantity (in expectation over words and contexts) is invariant to the choice of $\mathbf{t}$, so we can remove it from our minimization objective.

Now using the property that $\log \frac{a}{b} = -\log \frac{b}{a}$, we see that our minimization problem comes out to maximizing the HDMI:

$$\begin{aligned}\min_{\mathbf{t}} D_{\mathrm{KL}}(p_L(w_i|\mathbf{w}_{<i})||p_{\mathbf{t}}(w_i|t_i)) &= \min_{\mathbf{t}} -\mathbb{E}\left[\log \frac{p_{\mathbf{t}}(w_i|t_i)}{p(w_i)}\right] \\ &= \min_{\mathbf{t}} -I[W:T] \\ &= \max_{\mathbf{t}} I[W:T].\end{aligned}$$

□

Proposition 1 means that, if dependency structures are to be interpreted in the sense of Eq. 1, then heads and dependents will be those word pairs with maximal mutual information. This is our proposed theoretical justification for the HDMI Hypothesis.

5 Conclusion

We addressed the question of how syntactic dependency structure is reflected in the statistical covariance structure of words in natural language corpora, from an empirical and theoretical perspective. We advanced a theoretical argument, based on an information-theoretic interpretation of the postulates of dependency grammar, claiming that syntactic heads and dependents should correspond to word pairs with high MI: the HDMI Hypothesis. We reported what we believe is to date the largest-scale attempt to quantify mutual information between words in natural language text as a function of dependency structure, and found empirical support for the HDMI Hypothesis. We also found that MI between raw wordforms cannot be estimated by maximum likelihood estimation without bias even with 320 million tokens of text, and that MI between POS tags falls off with distance, mirroring previous findings about MI between orthographic letters (Li, 1989; Lin and Tegmark, 2017), although we found no fall-off for MI between distributional clusters.

Our work establishes a general link between syntactic structure and the statistical properties of texts, joining other work which has established connections between grammatical rules and information-theoretic statistics (Dębowski, 2015). We believe the HDMI Hypothesis can form the basis for improved grammar induction algorithms, by providing a new perspective on the head-outward generative models that have formed the basis of most work in that area. It also provides an intuitive means for comparatively evaluating different theories of dependency grammar (e.g., content-head vs. function-head: Osborne and Gerdes, 2019), in terms of the approximation error induced by different theories according to Eq. 3. In general, we believe the HDMI Hypothesis will also provide a stronger theoretical basis for corpus linguistics by linking the two conceptually independent notions of syntactic and statistical structure.

Acknowledgments

We thank Tim O'Donnell and Roger Levy for many discussions on these topics, and the anonymous reviewers for helpful comments on the paper.

References

Adger, D. (2018). The autonomy of syntax. In Hornstein, N., Lasnik, H., Patel-Grosz, P., and Yang, C., editors, *Syntactic Structures after 60 Years*, pages 153–175.

Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., and Collins, M. (2016). Globally normalized transition-based networks. In *Proceedings of the 54th Meeting of the Association for Computational Linguistics*, Berlin.

Archer, E., Park, I. M., and Pillow, J. W. (2013). Bayesian and quasi-Bayesian estimators for mutual information from discrete data. *Entropy*, 15(5):1738–1755.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188.

Bentz, C., Alikaniotis, D., Cysouw, M., and Ferrer-i-Cancho, R. (2017). The entropy of words—Learnability and expressivity across more than 1000 languages. *Entropy*, 19:275–307.

Buck, C., Heafield, K., and Van Ooyen, B. (2014). N-gram counts and language models from the common crawl. In *LREC*.

Chomsky, N. (1957). *Syntactic structures*. Walter de Gruyter.

Chow, C. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467.

Cover, T. M. and Thomas, J. (2006). *Elements of Information Theory*. John Wiley & Sons, Hoboken, NJ.

de Paiva Alves, E. (1996). The selection of the most probable dependency structure in Japanese using mutual information. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 372–374.

Dębowski, Ł. (2015). The relaxed Hilberg conjecture: A review and new experimental support. *Journal of Quantitative Linguistics*, 22(4):311–337.

Eisner, J. M. (1996). Three new probabilistic models for dependency parsing: An exploration. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 340–345.

Eisner, J. M. (1997). An empirical comparison of probability models for dependency grammar. Technical report, IRCS Report 96–11, University of Pennsylvania.

Futrell, R. and Levy, R. (2017). Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 688–698, Valencia, Spain.

Gulordava, K. and Merlo, P. (2015). Structural and lexical factors in adjective placement in complex noun phrases across romance languages. In *CoNLL*, pages 247–257.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(23):146–162.

Heim, I. and Kratzer, A. (1998). *Semantics in generative grammar*. Wiley-Blackwell, Malden, MA.

Hudson, R. A. (1984). *Word Grammar*. Blackwell.

Hudson, R. A. (2010). *An introduction to word grammar*. Cambridge University Press.

Jiang, Y., Han, W., and Tu, K. (2016). Unsupervised neural dependency parsing. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 763–771.

Klein, D. and Manning, C. D. (2004). Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, page 478.

Le, P. and Zuidema, W. (2015). Unsupervised dependency parsing: Let's use supervised parsers. *arXiv preprint arXiv:1504.04666*.

Li, W. (1989). Mutual information functions of natural language texts. Technical report, Santa Fe Institute Working Paper #1989-10-008.

Lin, H. W. and Tegmark, M. (2017). Critical behavior in physics and probabilistic formal languages. *Entropy*, 19(7):299.

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

Osborne, T. and Gerdes, K. (2019). The status of function words in dependency grammar: A critique of Universal Dependencies (UD). *Glossa: A Journal of General Linguistics*, 4(1):17.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253.

Pate, J. K. and Johnson, M. (2016). Grammar induction from (lots of) words alone. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 23–32.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Spitkovsky, V. I., Alshawi, H., and Jurafsky, D. (2012). Three dependency-and-boundary models for grammar induction. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*.

Yuret, D. (1998). *Discovery of linguistic relations using lexical attraction*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.