Variance of average surprisal: a better predictor for quality of grammar from unsupervised PCFG induction

Lifeng Jin and **William Schuler**
Department of Linguistics
The Ohio State University, Columbus, OH, USA
{jin, schuler}@ling.osu.edu

Abstract

In unsupervised grammar induction, data likelihood is known to be only weakly correlated with parsing accuracy, especially at convergence after multiple runs. In order to find a better indicator for quality of induced grammars, this paper correlates several linguistically- and psycholinguistically-motivated predictors to parsing accuracy on a large multilingual grammar induction evaluation data set. Results show that variance of average surprisal (VAS) better correlates with parsing accuracy than data likelihood, and that using VAS instead of data likelihood for model selection provides a significant accuracy boost. Further evidence shows VAS to be a better candidate than data likelihood for predicting word order typology classification. Analyses show that VAS seems to separate content words from function words in natural language grammars, and to better arrange words with different frequencies into separate classes that are more consistent with linguistic theory.

1 Introduction

Unsupervised grammar induction models learn to produce hierarchical structures for strings of words. Previous work (Seginer, 2007; Ponvert et al., 2011; Shain et al., 2016; Jin et al., 2018b) show that using data likelihood as both the objective for optimization and the criterion for model selection, either implicitly (in the case of Bayesian models) or explicitly (in the case of EM), gives good results on grammar induction. However, it is also known that data likelihood is only weakly correlated with parsing accuracy, especially at convergence (Smith, 2006; Johnson et al., 2007; Jin et al., 2018a). This weak correlation points to the fact that the maximization of data likelihood at convergence may be non-optimal for model selection, and this non-optimality indicates other constraints on learning may be at work in human acquisition. In this work, several linguistically- and psycholinguistically-motivated constraints related to syntax are explored as predictors of parsing accuracy for grammars learned by unsupervised induction (Jin et al., 2018a). Results show that variance of average surprisal (VAS) is better correlated with parsing accuracy of induced grammars than data likelihood. Using VAS for model selection at convergence also produces significantly higher parsing accuracy. Further evidence shows VAS to be a better candidate than data likelihood for predicting word order typology classification. Analyses show that VAS seems to separate content words from function words in natural language grammars, and seems to better arrange words with different frequencies into separate classes that are more consistent with linguistic theory.

2 Related work

Induction of PCFGs has previously been considered a difficult problem (Carroll and Charniak, 1992; Johnson et al., 2007; Liang et al., 2009; Tu, 2012). Earlier work attributed the lack of success for induction to a lack of correlation between parsing accuracy and data likelihood (Johnson et al., 2007), or to the likelihood function or the posterior being filled with weak local optima (Liang et al., 2009; Gimpel and Smith, 2012). Later work has shown that it is possible to induce PCFGs with useful labels from words alone (Shain et al., 2016; Jin et al., 2018b,a). Induction models of constituency grammars or trees usually use data likelihood as both the objective and the model selection criterion (Seginer, 2007; Johnson et al., 2007; Ponvert et al., 2011; Shen et al., 2018), but the weak correlation between data likelihood and parsing accuracy hints at the non-optimality of this practice (Smith, 2006; Headden et al., 2009; Jin

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2453–2463
Florence, Italy, July 28 - August 2, 2019.

et al., 2018a).

On the other hand, many linguistic and psycholinguistic theories propose constraints either as properties of natural language grammar or as constraints on human processing and acquisition. Chomsky (1965) proposes that grammars should favor fewer rules, which may be trimmed by the generalizability of the rules (Yang, 2017). Dryer (1992) argues that grammars with certain constituent ordering should produce trees with consistent branching tendencies, which is in contrast to theories that attribute constituent ordering to processing (Hawkins, 1994; Gibson, 1998). Rajkumar et al. (2016) and Jin et al. (2018b) show that grammars should generally control the maximal allowed stack depth. Yang (2013) observes that rules in a natural language grammar follow Zipf's law, just like words. Grammars may also contribute to the observation that the likelihood of each sentence tends to decrease as a monologue goes on (Keller, 2004; Levy and Jaeger, 2007).

3 Predictors

Motivated by these constraints, six accuracy predictors — data likelihood, right-branching score, rule complexity, average stack depth, Zipf likelihood ratio and variance of average surprisal — are evaluated as predictors of parsing accuracy over grammars from multiple runs of a PCFG inducer (Jin et al., 2018a). Variance of average surprisal, Zipf likelihood ratio and data likelihood are defined on the PCFG itself, and the other three are defined on Viterbi parses produced by the PCFG on the corpus.

Data likelihood

One of the most common induction and model selection criteria is data likelihood. Data likelihood (LL) refers to the marginal likelihood of a corpus given a PCFG, marginalizing out all trees:

$$\text{LL} = \mathsf{P}(\sigma; \mathbf{G}) = \sum_{\tau \in \mathcal{T}} \mathsf{P}(\sigma, \tau; \mathbf{G}), \qquad (1)$$

where σ is a corpus and $\mathcal{T}$ is all possible parse trees generated by a grammar $\mathbf{G}$ for σ. As it is usually the optimization objective, likelihood should be positively correlated with parsing accuracy at convergence.

Right-branching score

Branching Direction Theory (Dryer, 1992) explains different patterns of word order among languages. It distinguishes 'verb patterners,' which are non-phrasal lexical categories, from 'object patterners,' which are phrasal categories. It predicts that VO languages tend towards right-branching structures and OV languages tend towards left-branching structures. Let $|c_{\text{right}} \rightarrow a\ b|$ be the number of right children of a parent expanding into two non-terminal categories in all parse trees, and $|c_* \rightarrow a\ b|$ be the total number of nodes that expand into two non-terminal categories, then

$$\text{RBS} = \frac{|c_{\text{right}} \rightarrow a\ b|}{|c_* \rightarrow a\ b|} \qquad (2)$$

is the right branching score of the parse trees. A purely right-branching set of binary-branching trees yields an RBS of 1.0, and a purely left-branching set of binary-branching trees yields an RBS of 0.0. Previous work shows that right-branching baselines are accurate for a few languages (Seginer, 2007). BDT predicts that different word orders favor different branching directions.

Rule complexity

One of the evaluation metrics used in the generative linguistics tradition is the complexity of a grammar (Chomsky, 1965). Often the number of rules is used as a proxy measurement of how complex a proposed grammatical analysis is against some other reference grammatical analysis. According to this theory, fewer unique rules present in the Viterbi parses would indicate higher grammar quality.

Average stack depth

Embedding depth is a known limiting factor to human sentence processing (Chomsky and Miller, 1963; Wu, 2010; Rajkumar et al., 2016), and is shown to benefit unsupervised grammar induction (Noji and Johnson, 2016; Jin et al., 2018b). It is also evaluated in this work as a predictor of parsing accuracy, defined as the expected number of stack elements per sentence in a left-corner parser for the Viterbi parses. Theories such as that of Chomsky and Miller (1963) predict it to correlate negatively with parsing accuracy.

Zipf likelihood ratio

The distribution of words in a corpus is known to follow Zipf's law (Zipf, 1935), in which the frequency of a word is inversely proportional to its frequency rank. Counts of syntactic rules in annotated corpora also follow this law (Yang, 2013).

Motivated by this observation, experiments in this work also evaluate expected counts of all possible rules, and compute the ratio (Zipf R) between the likelihood that the rules are generated by a power law model and the likelihood that they are generated by a lognormal model of which the mean μ must be positive (Clauset et al., 2009). The higher the ratio, the better fit the power law model provides to the rule counts. Zipfian observations predict this ratio should be positively correlated with parsing accuracy.

Variance of average surprisal

Finally, languages may have other interesting properties that are not identified by maximizing the likelihood of the corpus. For example, languages often distinguish function words from content words and assign them distinct categories. If grammars assign very small sets of high frequency words to a few function-word-like categories, this will increase the difference in likelihood between sentences consisting of mostly these function words and sentences with more modifiers and other content words. The magnitude of this difference can be measured using variance of average sentential surprisal (VAS):

$$\text{VAS} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\log \mathsf{P}(\sigma_i)}{|\sigma_i|} - \frac{1}{N}\sum_{j=1}^{N}\frac{\log \mathsf{P}(\sigma_j)}{|\sigma_j|}\right)^2 \quad (3)$$

where N is the number of sentences in the corpus, and σ_i is the i-th sentence. Because sentences in larger corpora contain different numbers of function words, VAS is predicted to be high when the distinction between predicted function words and predicted content words in the induced grammar aligns with human judgments, indicating that VAS should be positively correlated with parsing accuracy.

4 Dataset

The grammar accuracy predictors described above are evaluated on multiple languages using corpora annotated with constituents (Xia et al., 2000; Marcus et al., 1993; Alastair et al., 2018) and corpora annotated with dependencies (Nivre et al., 2016) which are converted to constituents (Collins et al., 1999). An example is shown in Figure 1. These evaluations use corpora with at least 2,000 annotated sentences, excluding all sentences with non-projective dependency graphs.

Each induction run uses approximately 15,000 sentences randomly sampled from each language corpus. Languages with fewer than 15,000 annotated sentences are augmented with sentences sampled from Wikipedia (Zeman et al., 2017).

Evaluations initially screen predictors on a development partition consisting of 12 languages from 12 language subgroups covering language families including Indo-European, Uralic, Korean, Turkic, Sino-Tibetan and Afro-Asiatic. Significance tests use a separate test partition consisting of 25 languages[1] which are different from the development partition, covering additional Japanese, Austronesian and Austro-Asiatic language families.

5 Model

These evaluations use the Bayesian PCFG induction model from Jin et al. (2018a),[2] the objective function of which can be considered to be data likelihood.[3] However, the results for model selection reported in this paper are endemic neither to PCFG induction nor to the objective function used in induction. These experiments can be done with PCFGs randomly sampled from any distribution, but the fact that maximizing data likelihood as the objective can give better models than arbitrary random models ensures that evaluations are tractable and meaningful.

This model defines a Chomsky normal form (CNF) PCFG as a matrix $\mathbf{G}$ of binary rule probabilities which is first drawn from the Dirichlet prior with a concentration parameter β:

$$\mathbf{G} \sim \text{Dirichlet}(\beta) \quad (4)$$

Trees for sentences $1..N$ are then generated by drawing from a PCFG:

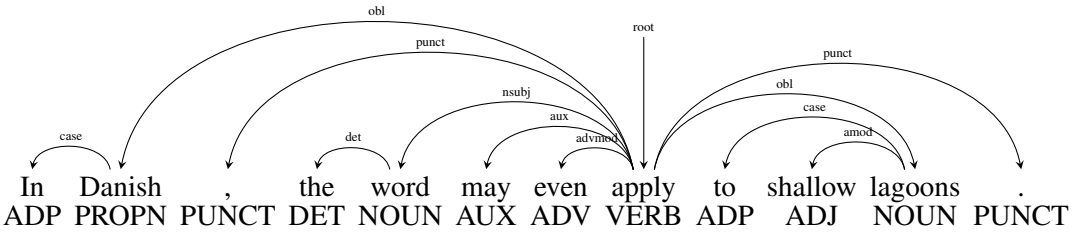
$$\tau_{1..N} \sim \text{PCFG}(\mathbf{G}) \quad (5)$$

Specifically, each tree τ is a set $\{\tau_\epsilon, \tau_1, \tau_2, \tau_{11}, \tau_{12}, \tau_{21}, ...\}$ of category node labels τ_η where $\eta \in \{1,2\}^*$ defines a path of left or right branches from the root to that node. Category labels for every pair of left and right children $\tau_{\eta 1}, \tau_{\eta 2}$ are drawn from a multinomial

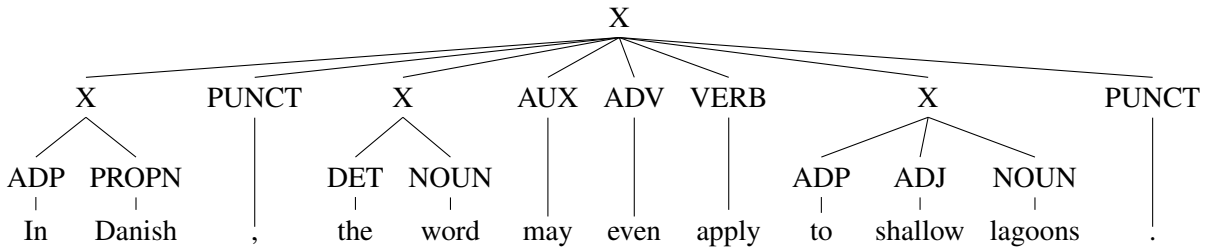
[1] Portuguese in the test partition refers to Brazilian Portuguese. Nynorsk and Bokmål are two varieties of Norwegian.

[2] https://github.com/lifengjin/dimi_emnlp18.

[3] Bayesian models usually have no objective function, but in inference the parameters will drift towards one of the modes, which may appear to be optimized for data likelihood.



(a) The dependency graph for the example sentence from the English Universal Dependency Treebank.



(b) The constituency tree converted from the dependency graph. Only the constituents where there is a single incoming dependency relation are kept. The three created constituents correspond to two PPs and one NP. They are labeled with X.

Figure 1: Examples of a dependency graph and the converted constituent tree for the sentence *In Danish, the word may even apply to shallow lagoons.*

distribution defined by the grammar $\mathbf{G}$ and the category of the parent τ_η:

$$\tau_{\eta 1}, \tau_{\eta 2} \sim \text{Multinomial}({\delta_{\tau_\eta}}^\top \mathbf{G}) \quad (6)$$

where δ_x is a Kronecker delta function equal to 1 at value x and 0 elsewhere, and terminals have null expansions $\mathsf{P}_\mathbf{G}(a\ b \mid w) = \mathsf{P}_\mathbf{G}(a\ b \mid \bot) = [\![a, b{=}\bot, \bot]\!]$ for $w \in W$.[4]

In inference, the conditional posteriors are calculated with a chart sampler (Johnson et al., 2007), and Gibbs sampling is used to draw samples of grammars and parse trees from the true posteriors. For example, at iteration t of Gibbs sampling:

$$\mathbf{G}^t \sim \mathsf{P}(\mathbf{G}^t \mid \tau_{1..N}^{t-1}, \sigma_{\tau_{1..N}^{t-1}}, \beta) \quad (7)$$

$$\tau_{1...N}^t \sim \mathsf{P}(\tau_{1..N}^t \mid \mathbf{G}^t, \sigma_{\tau_{1..N}^t}) \quad (8)$$

where σ_τ denotes the terminals in τ.

The inference procedure naturally produces sampled parses of a sentence, and the Viterbi parse of a sentence given an induced PCFG can be obtained by running the Viterbi algorithm with the grammar on the sentence.

6 Experiments

An exploratory evaluation on the 12-language development partition described in Section 4 measures the effectiveness of the proposed predictors in order to narrow the number of possible candidates prior to significance testing. A confirmatory evaluation on the 25-language test partition with significance testing is performed with the predictors that are found to be effective in the exploratory evaluation. Following Jin et al. (2018a), the concentration parameter of the Dirichlet priors is set to 0.2 for all languages. The number of syntactic categories C is set to 30 to allow the model to explore more complex syntactic structures. 30 random seeds are used for initialization of the model parameters, creating 30 runs for each language. The embedding depth of the induced grammars is not bounded in any run. All runs are stopped at iteration 700 which has been observed to have stable likelihood for at least 200 iterations (Jin et al., 2018a). A sampled grammar and Viterbi parse from the end of each run are used for predictor value calculation. Recall is used as the parsing accuracy metric for recovery of attested constituents.

7 Results

7.1 Development results

Correlation study

Columns two through seven in Table 1 show the correlation coefficients (Pearson's ρ) between all the proposed predictors and the recall of the Viterbi parses of the development partition. Coef-

[4] Here, $[\![\cdots]\!]$ is an indicator function.

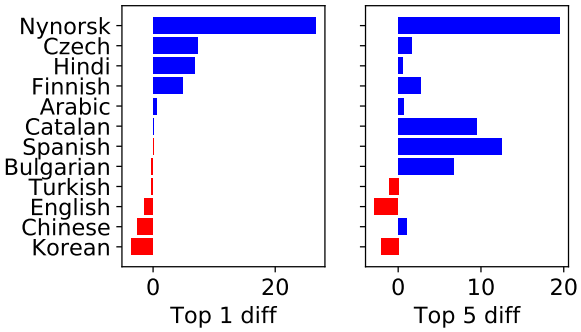


Figure 2: Recall difference between the run with the highest VAS and the highest likelihood as well as the difference between the average recall of the runs with the top 5 highest VAS and the top 5 highest likelihood on the development partition. Blue indicates that recall of the highest VAS runs is higher, and red indicates it is lower than the highest likelihood runs.

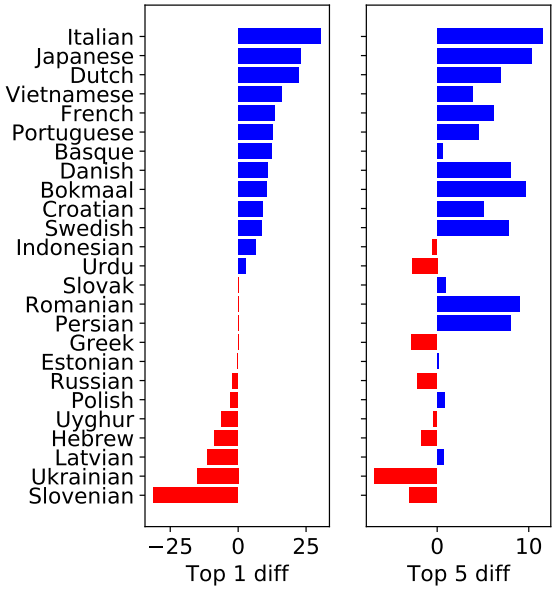


Figure 3: Recall difference between the run with the highest VAS and the highest likelihood as well as the difference between the average recall of the runs with the top 5 highest VAS and the top 5 highest likelihood on the test languages. Blue indicates that recall of the highest VAS runs is higher, and red indicates it is lower, than the highest likelihood runs.

ficients higher than 0.45 or lower than –0.45 are considered substantially predictive and reported in the table. Coefficients are averaged across reported languages.

Variance of average surprisal (VAS) has the highest correlation coefficients among all the predictors with the highest average correlation coefficient of 0.627. Data likelihood (LL), which is the most common metric for optimization and model selection in grammar induction, is the second best predictor. It also has a high average correlation coefficient of 0.588.[5]

Right-branching score also is substantially predictive of recall, but two of the languages have a negative coefficient, making it difficult to use as a model selection criterion without prior knowledge about the branching tendency of a language. Rule complexity, average stack depth as well as Zipf likelihood ratio all show up as predictive, but the signs of the coefficients are similarly inconsistent. Also, the signs of rule complexity are mostly positive, indicating that grammars should maintain a certain minimum level of complexity.

Parsing accuracy and model selection

The rightmost columns in Table 1 show parsing results on the development partition. The oracle recall is the highest recall obtained with 30 runs and the baseline reports whichever one of the left-branching baseline or the right-branching baseline has the highest recall, marked by L or R.

The VAS and LL columns in Table 1 show the parsing accuracy of the runs chosen by VAS and likelihood and Figure 2 shows the difference in recall. Positive difference shows that the run chosen with VAS is more accurate, and negative difference shows that LL is more accurate. Using VAS as the model selection criterion provides on average 3.19 points of recall gain. Recall gain from Nynorsk seems to be a fairly large outlier, but the positive gains from other languages are also larger than the negative gains. Figure 2 also shows the difference of average recall between the runs with the top 5 highest VAS and likelihood. There are still larger positive differences than negative differences, suggesting that VAS more strongly correlates with recall.

7.2 Test results

Parsing accuracy and model selection

In order to reduce the need for multiple trials correction, evaluations on the test partition only examine surprisal variance and data likelihood.

The VAS and LL columns in Table 2 show the parsing accuracy of the runs chosen by VAS and likelihood on the test partition, and Figure 3 shows

[5]Correlation coefficients using Kendall's τ are similar: on the development partition, the average τ is 0.27 for likelihood and 0.33 for VAS. On the test partition the average τ is 0.07 for likelihood and 0.24 for VAS.

Language	Correlation coefficients						Recall			
	Zipf R	Stack depth	RBS	Rule comp	LL	VAS	Baseline	LL	VAS	Oracle
Arabic	-	-	0.604	-	0.499	0.559	43.94 R	50.84	**51.39**	57.35
Bulgarian	–0.807	-	-	-	-	0.722	55.28 R	**70.65**	70.46	70.65
Catalan	–0.772	-	0.603	-	0.608	0.770	41.13 R	63.09	**63.20**	63.48
Chinese	-	-	-	-	-	0.532	29.19 R	**42.39**	39.88	42.39
Czech	-	-	-	–0.517	0.605	0.503	50.26 R	55.63	**62.88**	62.88
English	-	–0.540	0.554	0.549	0.689	0.673	44.74 R	**62.50**	61.11	65.57
Finnish	0.491	–0.700	0.854	-	-	-	**52.13 R**	46.27	51.16	54.16
Hindi	-	-	-	-	0.539	-	30.12 L	38.23	**45.10**	54.27
Korean	–0.545	0.868	–0.783	0.915	-	-	**40.38 R**	24.74	21.15	29.78
Nynorsk	-	-	0.576	-	-	0.677	55.40 R	41.46	**68.10**	68.20
Spanish	-	-	-	-	-	0.583	46.35 R	**53.83**	**53.83**	65.94
Turkish	–0.593	0.785	–0.954	0.512	-	-	**45.54 L**	33.94	33.61	47.02
Average	–0.445	0.103	0.207	0.365	0.588	0.627	44.54	48.63	**51.82**	56.81

Table 1: Correlation coefficients (Pearson's ρ) between recall at convergence and the proposed predictors on the languages in the development partition as well as recall from baselines and runs chosen with various model selection methods. Coefficients that are higher than 0.45 or lower than –0.45 are reported in table. Coefficients are averaged across reported languages. For recall, baseline shows recall from whichever one in left-branching baseline and right-branching baseline produces a higher recall. The direction of branching is marked by L or R. Oracle recall is from the oracle best run, and LL and VAS show recall from the run with the highest LL and highest VAS. The best run among the baseline, LL and VAS is boldfaced.

Language	Baseline	LL	VAS	Oracle
Basque	42.21 L	41.02	**53.31**	59.92
Bokmål	57.75 R	58.94	**69.28**	70.52
Croatian	47.43 R	50.97	**60.04**	60.04
Danish	55.30 R	58.91	**69.84**	69.84
Dutch	49.35 R	46.55	**68.73**	68.73
Estonian	48.08 R	**56.91**	56.71	56.91
French	42.22 R	47.25	**60.75**	63.09
Greek	49.62 R	**60.87**	56.41	64.66
Hebrew	43.52 R	**60.88**	**60.88**	65.20
Indonesian	50.37 R	50.90	**57.27**	57.27
Italian	52.98 R	38.39	**68.91**	70.61
Japanese	40.13 L	21.01	**44.04**	46.80
Latvian	51.67 R	**58.86**	47.67	58.86
Persian	24.40 R	**38.50**	**38.50**	42.22
Polish	70.33 R	**76.76**	73.89	78.27
Portuguese	45.32 R	51.41	**64.00**	65.31
Romanian	47.61 R	**61.48**	**61.48**	61.48
Russian	50.45 R	**61.78**	59.62	61.78
Slovak	64.83 R	**72.49**	**72.49**	72.78
Slovenian	54.54 R	**67.23**	36.02	69.35
Swedish	53.77 R	60.25	**68.92**	68.92
Ukrainian	51.88 R	**60.32**	45.19	60.32
Urdu	29.62 L	31.33	**34.11**	42.65
Uyghur	**45.77 L**	35.55	29.41	48.88
Vietnamese	55.41 R	43.55	**59.74**	59.74
Average	48.98	52.66	**56.69**	61.77

Table 2: Parsing accuracy for languages in the test partition. See the caption of Table 1 for the description of the columns.

the difference in recall for top 1 and top 5 runs. The patterns are similar to the ones on the development set. Using VAS as the model selection criterion with the top 1 runs provides on average 4.03 points of recall gain.

Table 3 shows correlation coefficients for LL and VAS on languages in the test partition. Again the observed pattern is similar, if not more extreme, to what is seen on the development partition. The magnitude of the coefficients is consistent with findings in the development partition. Except for Basque, the sign for VAS-recall correlation is consistently positive, confirming that it is reliable to use VAS for model selection.

Confirmatory significance testing is performed on two sets of 25,000 randomly sampled parses from the runs with highest likelihood and highest VAS on all test languages. The parses are randomly permuted between the two sets, and the difference in recall between the two sets is measured. This permutation test shows that the average 4.03 recall gain in Table 2 is highly unlikely to be due to chance ($p < 0.0001$), showing that VAS produces significantly more accurate grammars in model selection than using likelihood.

7.3 Word-order typology prediction

If VAS is much more highly correlated to parsing accuracy than previous predictors, it is possible to use it as an unsupervised proxy to parsing accuracy. Branching Decision Theory (Dryer, 1992) predicts that VO languages favor right-branching structures and OV languages favor left-branching structures. This prediction can be evaluated by correlating VAS and RBS, and using the sign of the correlation coefficient as the word-order pre-

Lang.	LL	VAS	Lang.	LL	VAS
Basque	-	–0.578	Latvian	-	-
Bokmål	-	0.603	Persian	-	0.462
Croatian	-	0.615	Polish	-	-
Danish	-	0.551	Portuguese	-	0.484
Dutch	-	0.740	Romanian	-	0.644
Estonian	0.698	0.686	Russian	-	0.682
French	-	0.715	Slovak	-	0.522
Greek	-	0.452	Slovenian	-	-
Hebrew	0.600	0.667	Swedish	-	0.803
Indonesian	-	-	Ukrainian	-	-
Italian	-	0.481	Urdu	-	-
Japanese	-	0.627	Uyghur	-	-
Vietnamese	–0.458	-			
Average	0.280	0.539			

Table 3: Correlation coefficients between recall at convergence and the proposed predictors on the test partition. See the caption of Table 1 for the description of the columns.

diction. This tests if grammars following the branching tendency predicted by the theory should have higher parsing accuracy. Table 4 shows results for the VAS-RBS correlation reported along with a few baselines, including a uniform baseline, a majority baseline (where there is oracle knowledge about the data set that the majority of languages is VO), the LL-RBS correlation baseline (where data likelihood is used as the proxy for recall), as well as the recall-RBS oracle performance.

There are 29 VO languages and 7 OV languages in the data set (Dryer, 2011).[6] Macro F1 is reported for all systems here as the population distribution of OV and VO languages in the world is almost uniform (Dryer, 1992). First, as predicted by BDT, using signs of the correlation between recall and right-branching score yields the best macro F1 score. Second, using VAS as a proxy of recall yields a much higher F score than all the other baselines, including likelihood. In fact, likelihood performs the worst of all the baselines. This result shows again that the correlation between VAS and parsing accuracy is stronger than likelihood at convergence, and this tighter correlation can be useful in other unsupervised tasks.

8 Discussion

Positive effects for predictors other than data likelihood suggest that natural language grammars are not optimally learned to explain sentence forms, but may additionally reflect biological constraints

[6]Dutch has no dominant VO-OV order.

Model	Gold VO		Gold OV		Macro-f
	Right	Wrong	Right	Wrong	
Uniform	14.5	14.5	3.5	3.5	44.5
Majority	29	0	0	7	44.8
LL	11	18	5	2	42.9
VAS	19	10	7	0	**69.2**
Recall	*27*	*2*	*6*	*1*	*87.4*

Table 4: The macro-F1 scores for the task of predicting the word order of a language.

on grammar learning. In particular, the success of VAS may point to a bias toward a function/content distinction in natural language grammars, with common words more likely to form distinctive categories in human learners than co-occurrence statistics would suggest. This bias would produce the observed result that sentences containing more function words have higher per-word probabilities than sentences containing more content words and the existence of such a distinction may give rise to higher surprisal variance. In contrast, a lack of such bias would allow common words to mix with rare words, yielding more uniform probabilities and low surprisal variance, contrary to observations of conditions under which recall is maximized. The fact that simple maximization of data likelihood appears to favor the more uniform response suggests it is not a sufficient model of grammar learning.

We first evaluate this hypothesis by examining the ratio between content and function words across sentences to determine whether this ratio is constant in a language. We use the Wall Street Journal portion of the Penn Treebank as the target corpus,[7] and calculate the ratio of function to content words in all sentences, and examine the density of the ratio in terms of sentence count and its relationship with sentence length. The left figure in Figure 4 shows the relation between the function-content word ratio and sentence count. The function-content word ratio has a mode at around 0.7, but the count pass is also widely distributed mostly within the range between 0 and 1. This shows that the ratio between content and function words in a language does not appear to be constant. The right figure in Figure 4 shows the relationship between the function-content word ra-

[7]We consider words with part-of-speech tags like CC, DT, IN, MD, PDT, RP, TO, PRP, PRP$, WDT, WP, WP$, WRB and UH as function words, and words with POS tags like JJ, JJR, JJS, NN, NNS, NNP, NNPS, RB, RBR, RBS, VB, VBD, VBG, VBN, VBP, VBZ and FW as content words.

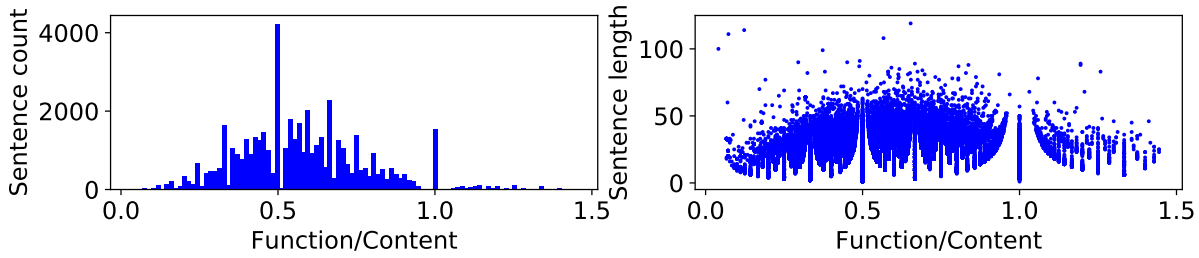


Figure 4: Left: the relationship between sentence count and the ratio between content and function words. Right: the relationship between sentence length and the ratio in the Wall Street Journal part of the Penn Treebank.

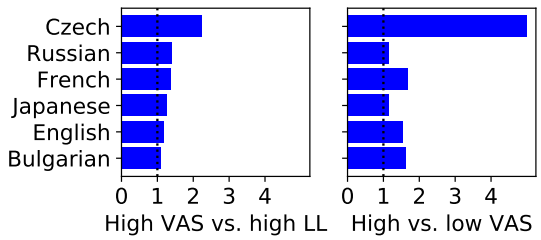


Figure 5: Left: Ratio of number of high joint probability words in the grammars from runs with highest VAS vs. the highest likelihood. Right: Ratio of number of high joint probability words in the grammars from runs with highest VAS vs. the lowest VAS.

tio to sentence length. The ratio seems to converge to 0.7 as the sentence gets longer, but the majority of the sentences in the corpus are below 50 words, and the spread of function-content word ratio for sentences with shorter lengths is also very wide.

In many languages, the words with highest frequencies are usually closed class words, such as prepositions and determiners, and these words typically split away from other major classes and form their own classes, raising their probabilities. Low frequency words, on the other hand, tend to move from smaller classes into larger classes, and thus lower their probabilities. It is known that low frequency words, especially hapax legomena, are usually open class words like nouns or adjectives. To reassign these words into larger classes may help them find a natural home where the majority is of the same class as the rare words. This strategy helps better assign words to syntactic classes, which in turn helps create syntactic rules which better align with human annotations.

The claim that VAS promotes a distinction between function and content words can be evaluated by comparing joint probabilities of the most frequent words in each language and their most common class in grammars from runs with highest VAS, lowest VAS and highest likelihood. In each case, if the most frequent words have higher probabilities in the high VAS run, this may suggest VAS is correlated with function-content distinctions. Figure 5 shows the top 50 most frequent words in 6 different languages with substantial correlations between VAS and recall.

The left figure shows the fraction of words in the run with the highest VAS that have joint probabilities of words and their generating categories higher than in the run with the highest likelihood (i.e. words that have higher probabilities in VAS-selected grammars than likelihood-selected grammars). The right figure shows the fraction of words in the run with the highest VAS that have joint probabilities higher than in the run with the lowest VAS (i.e. words that have higher probabilities in VAS-selected grammars than in VAS-dispreferred grammars). For all six languages, the ratio of words with higher joint probability is larger than 1, meaning that frequent words in the run with the highest VAS are assigned to classes with higher joint probabilities than words in the run with the highest likelihood or the run with the lowest VAS, consistent with the hypothesis that VAS promotes a distinction between function and content words. Probabilities for some example words are shown in Figure 6.

A different explanation may be considered that information content in a sentence is higher when the sentence is longer (Keller, 2004), and when VAS is maximized, grammars that produce uniform information content across different sentence length are disfavored. For example, punctuation contributes more to the likelihood of short sentences than to long sentences. Assigning high probabilities to punctuation may create the result of sentence likelihood co-varying with sentence length. For a grammar to conform to this rule may help it produce structures more in line with hu-

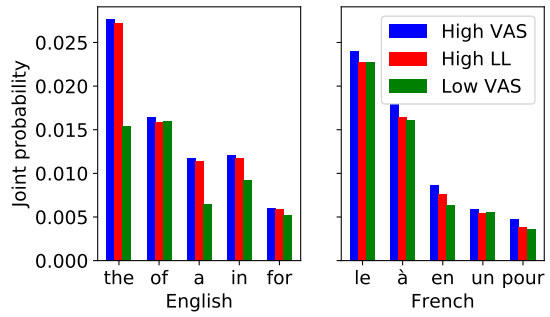


Figure 6: Example high frequency words from the highest VAS, the highest likelihood and the lowest VAS runs in English and French.

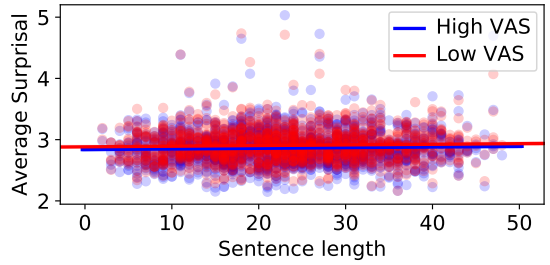


Figure 7: The distribution of VAS values across sentences of different lengths in the highest VAS run and the lowest VAS run for English. The correlations between VAS and sentence length in both runs are insignificant.

man annotations in the data set. Figure 7 shows the distribution of VAS plotted against sentence length. The regression lines for both the highest VAS and lowest VAS cases show a flat slope indicating the correlation between VAS and sentence length is not substantial, which is supported by correlation testing with Kendall's τ test between sentence length and VAS in the high VAS run ($\tau = -0.01, p = 0.41$) and in the low VAS run ($\tau = -0.02, p = 0.28$). This shows that the effectiveness of VAS cannot be explained by the hypothesis that it guides the grammar to generate syntactic structures by shaping the sentential information content to co-vary with sentence length.

9 Conclusion

This work explores the non-optimality of data likelihood for model selection in unsupervised grammar induction. Experiments with several linguistically- and psycholinguistically-motivated predictors on a large multilingual data set show that variance of average surprisal (VAS) is highly predictive of parsing performance. Using it as the criterion for model selection outperforms data likelihood significantly. Further evidence shows VAS to be a better candidate than data likelihood for predicting word-order typology. Analyses show that VAS seems to separate content words from function words in natural language grammars and better arrange words with different frequencies into different classes that are more consistent with these linguistic distinctions.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful comments. Computations for this project were partly run on the Ohio Supercomputer Center (1987). This research was partially funded by the Defense Advanced Research Projects Agency award HR0011-15-2-0022. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. This work was also supported by the National Science Foundation grant 1816891. All views expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Butler Alastair, Kei Yoshimoto, Shota Hiyama, Stephen Wright Horn, Iku Nagasaki, and Ai Kubota. 2018. The Keyaki Treebank Parsed Corpus.

Glenn Carroll and Eugene Charniak. 1992. Two experiments on learning probabilistic dependency grammars from corpora. *Working Notes of the Workshop on Statistically-Based NLP Techniques*, (March):1–13.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Noam Chomsky and George A Miller. 1963. Introduction to the formal analysis of natural languages. In *Handbook of Mathematical Psychology*, pages 269–321. Wiley, New York, NY.

Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. 2009. Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703.

Michael Collins, Lance Ramshaw, Jan Hajič, and Christoph Tillmann. 1999. A Statistical Parser for Czech. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 505–512.

Matthew S Dryer. 1992. The Greenbergian Word Order Correlations. *Language*, 68(1):81–138.

Matthew S Dryer. 2011. The evidence for word order correlations. *Linguistic Typology*, 15(2):335–380.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Kevin Gimpel and Noah A Smith. 2012. Concavity and Initialization for Unsupervised Dependency Parsing. In *NAACL*, pages 577–581.

John A Hawkins. 1994. *A performance theory of order and constituency*. Cambridge University Press, Cambridge, U.K.

William P. Headden, III, Mark Johnson, and David McClosky. 2009. Improving unsupervised dependency parsing with richer contexts and smoothing. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 101–109.

Lifeng Jin, Finale Doshi-Velez, Timothy A Miller, William Schuler, and Lane Schwartz. 2018a. Depth-bounding is effective: Improvements and evaluation of unsupervised PCFG induction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Lifeng Jin, Finale Doshi-Velez, Timothy A Miller, William Schuler, and Lane Schwartz. 2018b. Unsupervised Grammar Induction with Depth-bounded PCFG. *Transactions of the Association for Computational Linguistics*.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007. Bayesian Inference for PCFGs via Markov chain Monte Carlo. *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 139–146.

Frank Keller. 2004. The Entropy Rate Principle as a Predictor of Processing Effort : An Evaluation against Eye-tracking Data. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 317–324.

Roger Levy and Florian T. Jaeger. 2007. Speakers optimize information density through syntactic reduction. *Advances in Neural Information Processing Systems*.

Percy Liang, Michael I Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - ACL-IJCNLP '09*, volume 1, page 91.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D Manning, Ryan Mcdonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of Language Resources and Evaluation Conference*.

Hiroshi Noji and Mark Johnson. 2016. Using Left-corner Parsing to Encode Universal Structural Constraints in Grammar Induction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 33–43.

The Ohio Supercomputer Center. 1987. Ohio Supercomputer Center. \url{http://osc.edu/ark:/19495/f5s1ph73}.

Elias Ponvert, Jason Baldridge, and Katrin Erk. 2011. Simple unsupervised grammar induction from raw text with cascaded finite state models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1077–1086.

Rajakrishnan Rajkumar, Marten Van Schijndel, Michael White, and William Schuler. 2016. Investigating locality effects and surprisal in written English syntactic choice phenomena. *Cognition*, 155:204–232.

Yoav Seginer. 2007. Fast Unsupervised Incremental Parsing. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, pages 384–391.

Cory Shain, William Bryce, Lifeng Jin, Victoria Krakovna, Finale Doshi-Velez, Timothy Miller, William Schuler, and Lane Schwartz. 2016. Memory-bounded left-corner unsupervised grammar induction on child-directed input. In *Proceedings of the International Conference on Computational Linguistics*, pages 964–975.

Yikang Shen, Zhouhan Lin, Chin-Wei Huang, and Aaron Courville. 2018. Neural Language Modeling by Jointly Learning Syntax and Lexicon. In *ICLR*.

Noah Ashton Smith. 2006. Novel Estimation Methods for Unsupervised Discovery of Latent Structure in Natural Language Text. *PhD Thesis*, pages 1–228.

Kewei Tu. 2012. *Unsupervised learning of probabilistic grammars*. Ph.D. thesis.

Stephen Wu. 2010. Complexity Metrics in an Incremental Right-corner Parser. In *Proceedings of the North American Association for Computational Linguistics*.

Fei Xia, Martha Palmer, Nianwen Xue, Mary Ellen Ocurowski, John Kovarik, Fu-Dong Chiou, Shizhe Huang, Tony Kroch, and Mitch Marcus. 2000. Developing Guidelines and Ensuring Consistency for Chinese Text Annotation. In *Proceedings of the Second Language Resources and Evaluation Conference*.

Charles Yang. 2013. Who s Afraid of George Kingsley Zipf ? *Significance*, 10(6):29–34.

Charles Yang. 2017. Rage against the machine: Evaluation metrics in the 21st century. *Language Acquisition*, 24(2):100–125.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Vclava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria DePaiva, Kira Droganova, Hctor Martínez Alonso, ar Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, volume 17, pages 1–19.

George K. Zipf. 1935. *The Psychobiology of Language*. Houghton-Mifflin.