# Using Universal Linguistic Knowledge to Guide Grammar Induction

**Tahira Naseem, Harr Chen, Regina Barzilay**
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
{tahira, harr, regina} @csail.mit.edu

**Mark Johnson**
Department of Computing
Macquarie University
mark.johnson@mq.edu.au

## Abstract

We present an approach to grammar induction that utilizes syntactic universals to improve dependency parsing across a range of languages. Our method uses a single set of manually-specified language-independent rules that identify syntactic dependencies between pairs of syntactic categories that commonly occur across languages. During inference of the probabilistic model, we use posterior expectation constraints to require that a minimum proportion of the dependencies we infer be instances of these rules. We also automatically refine the syntactic categories given in our coarsely tagged input. Across six languages our approach outperforms state-of-the-art unsupervised methods by a significant margin.[1]

## 1 Introduction

Despite surface differences, human languages exhibit striking similarities in many fundamental aspects of syntactic structure. These structural correspondences, referred to as *syntactic universals*, have been extensively studied in linguistics (Baker, 2001; Carnie, 2002; White, 2003; Newmeyer, 2005) and underlie many approaches in multilingual parsing. In fact, much recent work has demonstrated that learning cross-lingual correspondences from corpus data greatly reduces the ambiguity inherent in syntactic analysis (Kuhn, 2004; Burkett and Klein, 2008; Cohen and Smith, 2009a; Snyder et al., 2009; Berg-Kirkpatrick and Klein, 2010).

[1]The source code for the work presented in this paper is available at http://groups.csail.mit.edu/rbg/code/dependency/

| | |
|---|---|
| Root → Auxiliary | Noun → Adjective |
| Root → Verb | Noun → Article |
| Verb → Noun | Noun → Noun |
| Verb → Pronoun | Noun → Numeral |
| Verb → Adverb | Preposition → Noun |
| Verb → Verb | Adjective → Adverb |
| Auxiliary → Verb | |

Table 1: The manually-specified universal dependency rules used in our experiments. These rules specify head-dependent relationships between coarse (i.e., unsplit) syntactic categories. An explanation of the ruleset is provided in Section 5.

In this paper, we present an alternative grammar induction approach that exploits these structural correspondences by declaratively encoding a small set of universal dependency rules. As input to the model, we assume a corpus annotated with coarse syntactic categories (i.e., high-level part-of-speech tags) and a set of universal rules defined over these categories, such as those in Table 1. These rules incorporate the definitional properties of syntactic categories in terms of their interdependencies and thus are universal across languages. They can potentially help disambiguate structural ambiguities that are difficult to learn from data alone — for example, our rules prefer analyses in which verbs are dependents of auxiliaries, even though analyzing auxiliaries as dependents of verbs is also consistent with the data. Leveraging these universal rules has the potential to improve parsing performance for a large number of human languages; this is particularly relevant to the processing of low-resource

languages. Furthermore, these universal rules are compact and well-understood, making them easy to manually construct.

In addition to these universal dependencies, each specific language typically possesses its own idiosyncratic set of dependencies. We address this challenge by requiring the universal constraints to only hold in expectation rather than absolutely, i.e., we permit a certain number of violations of the constraints.

We formulate a generative Bayesian model that explains the observed data while accounting for declarative linguistic rules during inference. These rules are used as expectation constraints on the posterior distribution over dependency structures. This approach is based on the posterior regularization technique (Graça et al., 2009), which we apply to a variational inference algorithm for our parsing model. Our model can also optionally refine common high-level syntactic categories into per-language categories by inducing a clustering of words using Dirichlet Processes (Ferguson, 1973). Since the universals guide induction toward linguistically plausible structures, automatic refinement becomes feasible even in the absence of manually annotated syntactic trees.

We test the effectiveness of our grammar induction model on six Indo-European languages from three language groups: English, Danish, Portuguese, Slovene, Spanish, and Swedish. Though these languages share a high-level Indo-European ancestry, they cover a diverse range of syntactic phenomenon. Our results demonstrate that universal rules greatly improve the accuracy of dependency parsing across all of these languages, outperforming current state-of-the-art unsupervised grammar induction methods (Headden III et al., 2009; Berg-Kirkpatrick and Klein, 2010).

## 2 Related Work

**Learning with Linguistic Constraints** Our work is situated within a broader class of unsupervised approaches that employ declarative knowledge to improve learning of linguistic structure (Haghighi and Klein, 2006; Chang et al., 2007; Graça et al., 2007; Cohen and Smith, 2009b; Druck et al., 2009; Liang et al., 2009a). The way we apply constraints is closest to the latter two approaches of posterior regularization and generalized expectation criteria.

In the posterior regularization framework, constraints are expressed in the form of expectations on posteriors (Graça et al., 2007; Ganchev et al., 2009; Graça et al., 2009; Ganchev et al., 2010). This design enables the model to reflect constraints that are difficult to encode via the model structure or as priors on its parameters. In their approach, parameters are estimated using a modified EM algorithm, where the E-step minimizes the KL-divergence between the model posterior and the set of distributions that satisfies the constraints. Our approach also expresses constraints as expectations on the posterior; we utilize the machinery of their framework within a variational inference algorithm with a mean field approximation.

Generalized expectation criteria, another technique for declaratively specifying expectation constraints, has previously been successfully applied to the task of dependency parsing (Druck et al., 2009). This objective expresses constraints in the form of preferences over model expectations. The objective is penalized by the square distance between model expectations and the prespecified values of the expectation. This approach yields significant gains compared to a fully unsupervised counterpart. The constraints they studied are corpus- and language-specific. Our work demonstrates that a small set of language-independent universals can also serve as effective constraints. Furthermore, we find that our method outperforms the generalized expectation approach using corpus-specific constraints.

**Learning to Refine Syntactic Categories** Recent research has demonstrated the usefulness of automatically refining the granularity of syntactic categories. While most of the existing approaches are implemented in the supervised setting (Finkel et al., 2007; Petrov and Klein, 2007), Liang et al. (2007) propose a non-parametric Bayesian model that learns the granularity of PCFG categories in an unsupervised fashion. For each non-terminal grammar symbol, the model posits a Hierarchical Dirichlet Process over its refinements (subsymbols) to automatically learn the granularity of syntactic categories. As with their work, we also use non-parametric priors for category refinement and em-

ploy variational methods for inference. However, our goal is to apply category refinement to dependency parsing, rather than to PCFGs, requiring a substantially different model formulation. While Liang et al. (2007) demonstrated empirical gains on a synthetic corpus, our experiments focus on unsupervised category refinement on real language data.

**Universal Rules in NLP** Despite the recent surge of interest in multilingual learning (Kuhn, 2004; Cohen and Smith, 2009a; Snyder et al., 2009; Berg-Kirkpatrick and Klein, 2010), there is surprisingly little computational work on linguistic universals. On the acquisition side, Daumé III and Campbell (2007) proposed a computational technique for discovering universal implications in typological features. More closely related to our work is the position paper by Bender (2009), which advocates the use of manually-encoded cross-lingual generalizations for the development of NLP systems. She argues that a system employing such knowledge could be easily adapted to a particular language by specializing this high level knowledge based on the typological features of the language. We also argue that cross-language universals are beneficial for automatic language processing; however, our focus is on learning language-specific adaptations of these rules from data.

## 3 Model

The central hypothesis of this work is that unsupervised dependency grammar induction can be improved using universal linguistic knowledge. Toward this end our approach is comprised of two components: a probabilistic model that explains how sentences are generated from latent dependency structures and a technique for incorporating declarative rules into the inference process.

We first describe the generative story in this section before turning to how constraints are applied during inference in Section 4. Our model takes as input (i.e., as observed) a set of sentences where each word is annotated with a coarse part-of-speech tag. Table 2 provides a detailed technical description of our model's generative process, and Figure 1 presents a model diagram.

For each observed coarse symbol $s$:

1. Draw top-level infinite multinomial over subsymbols $\beta_s \sim \mathrm{GEM}(\gamma)$.
2. For each subsymbol $z$ of symbol $s$:
   (a) Draw word emission multinomial $\phi_{sz} \sim \mathrm{Dir}(\phi_0)$.
   (b) For each context value $c$:
      i. Draw child symbol generation multinomial $\theta_{szc} \sim \mathrm{Dir}(\theta_0)$.
      ii. For each child symbol $s'$:
         A. Draw second-level infinite multinomial over subsymbols $\pi_{s'szc} \sim \mathrm{DP}(\alpha, \beta_{s'})$.

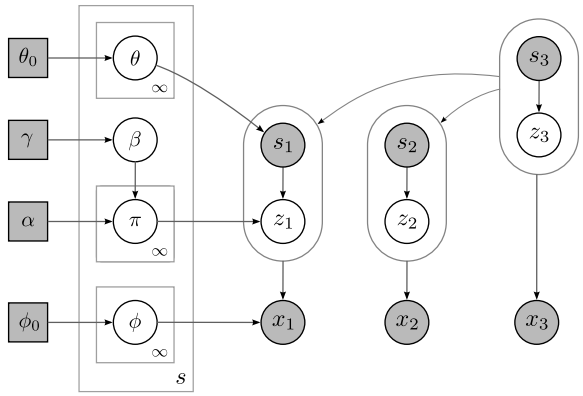
For each tree node $i$ generated in context $c$ by parent symbol $s'$ and parent subsymbol $z'$:

1. Draw coarse symbol $s_i \sim \mathrm{Mult}(\theta_{s'z'})$.
2. Draw subsymbol $z_i \sim \mathrm{Mult}(\pi_{s_i s' z' c})$.
3. Draw word $x_i \sim \mathrm{Mult}(\phi_{s_i z_i})$.

Table 2: The generative process for model parameters and parses. In the above GEM, DP, Dir, and Mult refer respectively to the stick breaking distribution, Dirichlet process, Dirichlet distribution, and multinomial distribution.

**Generating Symbols and Words** We describe how a single node of the tree is generated before discussing how the entire tree structure is formed. Each node of the dependency tree is comprised of three random variables: an observed coarse symbol $s$, a hidden refined subsymbol $z$, and an observed word $x$. In the following let the parent of the current node have symbol $s'$ and subsymbol $z'$; the root node is generated from separate root-specific distributions. Subsymbol refinement is an optional component of the full model and can be omitted by deterministically equating $s$ and $z$. As we explain at the end of this section, without this aspect the generative story closely resembles the classic dependency model with valence (DMV) of Klein and Manning (2004).

First we draw symbol $s$ from a finite multinomial




| | | |
|---|---|---|
| $s$ | - | coarse symbol (observed) |
| $z$ | - | refined subsymbol |
| $x$ | - | word (observed) |
| $\theta_{szc}$ | - | distr over child coarse symbols for each parent $s$ and $z$ and context $c$ |
| $\beta_s$ | - | top-level distr over subsymbols for $s$ |
| $\pi_{ss'z'c}$ | - | distr over subsymbols for each $s$, parent $s'$ and $z'$, and context $c$ |
| $\phi_{sz}$ | - | distr over words for $s$ and $z$ |

Figure 1: Graphical representation of the model and a summary of the notation. There is a copy of the outer plate for each distinct symbol in the observed coarse tags. Here, node 3 is shown to be the parent of nodes 1 and 2. Shaded variables are observed, square variables are hyperparameters. The elongated oval around $s$ and $z$ represents the two variables jointly. For clarity the diagram omits some arrows from $\theta$ to each $s$, $\pi$ to each $z$, and $\phi$ to each $x$.

distribution with parameters $\theta_{s'z'c}$. As the indices indicate, we have one such set of multinomial parameters for every combination of parent symbol $s'$ and subsymbol $z'$ along with a *context* $c$. Here the context of the current node can take one of six values corresponding to every combination of direction (left or right) and valence (first, second, or third or higher child) with respect to its parent. The prior (base distribution) for each $\theta_{s'z'c}$ is a symmetric Dirichlet with hyperparameter $\theta_0$.

Next we draw the refined syntactic category subsymbol $z$ from an infinite multinomial with parameters $\pi_{ss'z'c}$. Here the selection of $\pi$ is indexed by the current node's coarse symbol $s$, the symbol $s'$ and subsymbol $z'$ of the parent node, and the context $c$ of the current node.

For each unique coarse symbol $s$ we tie together the distributions $\pi_{ss'z'c}$ for all possible parent and context combinations (i.e., $s'$, $z'$, and $c$) using a Hierarchical Dirichlet Process (HDP). Specifically, for a single $s$, each distribution $\pi_{ss'z'c}$ over subsymbols is drawn from a DP with concentration parameter $\alpha$ and base distribution $\beta_s$ over subsymbols. This base distribution $\beta_s$ is itself drawn from a GEM prior with concentration parameter $\gamma$. By formulating the generation of $z$ as an HDP, we can share parameters for a single coarse symbol's subsymbol distribution while allowing for individual variability based on node parent and context. Note that parameters are not shared across different coarse symbols, preserving the distinctions expressed via the coarse tag annotations.

Finally, we generate the word $x$ from a finite multinomial with parameters $\phi_{sz}$, where $s$ and $z$ are the symbol and subsymbol of the current node. The $\phi$ distributions are drawn from a symmetric Dirichlet prior.

**Generating the Tree Structure** We now consider how the structure of the tree arises. We follow an approach similar to the widely-referenced DMV model (Klein and Manning, 2004), which forms the basis of the current state-of-the-art unsupervised grammar induction model (Headden III et al., 2009). After a node is drawn we generate children on each side until we produce a designated STOP symbol. We encode more detailed valence information than Klein and Manning (2004) and condition child generation on parent valence. Specifically, after drawing a node we first decide whether to proceed to generate a child or to stop conditioned on the parent symbol and subsymbol and the current context (direction and valence). If we decide to generate a child we follow the previously described process for constructing a node. We can combine the stopping decision with the generation of the child symbol by including a distinguished STOP symbol as a possible outcome in distribution $\theta$.

**No-Split Model Variant** In the absence of subsymbol refinement (i.e., when subsymbol $z$ is set to be identical to coarse symbol $s$), our model simplifies in some respects. In particular, the HDP gener-

ation of $z$ is obviated and word $x$ is drawn from a word distribution $\phi_s$ indexed solely by coarse symbol $s$. The resulting simplified model closely resembles DMV (Klein and Manning, 2004), except that it 1) explicitly generate words $x$ rather than only part-of-speech tags $s$, 2) encodes richer context and valence information, and 3) imposes a Dirichlet prior on the symbol distribution $\theta$.

## 4 Inference with Constraints

We now describe how to augment our generative model of dependency structure with constraints derived from linguistic knowledge. Incorporating arbitrary linguistic rules directly in the generative story is challenging as it requires careful tuning of either the model structure or priors for each constraint. Instead, following the approach of Graça et al. (2007), we constrain the posterior to satisfy the rules in expectation during inference. This effectively biases the inference toward linguistically plausible settings.

In standard variational inference, an intractable true posterior is approximated by a distribution from a tractable set (Bishop, 2006). This tractable set typically makes stronger independence assumptions between model parameters than the model itself. To incorporate the constraints, we further restrict the set to only include distributions that satisfy the specified expectation constraints over hidden variables.

In general, for some given model, let $\theta$ denote the entire set of model parameters and $z$ and $x$ denote the hidden structure and observations respectively. We are interested in estimating the posterior $p(\theta, z \mid x)$. Variational inference transforms this problem into an optimization problem where we try to find a distribution $q(\theta, z)$ from a restricted set $\mathcal{Q}$ that minimizes the KL-divergence between $q(\theta, z)$ and $p(\theta, z \mid x)$:

$$\begin{aligned}&\mathrm{KL}(q(\theta, z) \parallel p(\theta, z \mid x)) \\ &= \int q(\theta, z) \log \frac{q(\theta, z)}{p(\theta, z, x)} d\theta dz + \log p(x).\end{aligned}$$

Rearranging the above yields:

$$\log p(x) = \mathrm{KL}(q(\theta, z) \parallel p(\theta, z \mid x)) + \mathcal{F},$$

where $\mathcal{F}$ is defined as

$$\mathcal{F} \equiv \int q(\theta, z) \log \frac{p(\theta, z, x)}{q(\theta, z)} d\theta dz. \quad (1)$$

Thus $\mathcal{F}$ is a lower bound on likelihood. Maximizing this lower bound is equivalent to minimizing the KL-divergence between $p(\theta, z \mid x)$ and $q(\theta, z)$. To make this maximization tractable we make a mean field assumption that $q$ belongs to a set $\mathcal{Q}$ of distributions that factorize as follows:

$$q(\theta, z) = q(\theta) q(z).$$

We further constrain $q$ to be from the subset of $\mathcal{Q}$ that satisfies the expectation constraint $E_q[f(z)] \leq b$ where $f$ is a deterministically computable function of the hidden structures. In our model, for example, $f$ counts the dependency edges that are an instance of one of the declaratively specified dependency rules, while $b$ is the proportion of the total dependencies that we expect should fulfill this constraint.[2]

With the mean field factorization and the expectation constraints in place, solving the maximization of $\mathcal{F}$ in (1) separately for each factor yields the following updates:

$$q(\theta) = \operatorname*{argmin}_{q(\theta)} \mathrm{KL}\left(q(\theta) \parallel q'(\theta)\right), \quad (2)$$

$$q(z) = \operatorname*{argmin}_{q(z)} \mathrm{KL}\left(q(z) \parallel q'(z)\right) \quad s.t. \quad E_{q(z)}[f(z)] \leq b, \quad (3)$$

where

$$q'(\theta) \propto \exp E_{q(z)}[\log p(\theta, z, x)], \quad (4)$$

$$q'(z) \propto \exp E_{q(\theta)}[\log p(\theta, z, x)]. \quad (5)$$

We can solve (2) by setting $q(\theta)$ to $q'(\theta)$ — since $q(z)$ is held fixed while updating $q(\theta)$, the expectation function of the constraint remains constant during this update. As shown by Graça et al. (2007), the update in (3) is a constrained optimization problem and can be solved by performing gradient search on its dual:

$$\operatorname*{argmin}_{\lambda} \lambda^\top b + \log \sum_z q'(z) \exp(-\lambda^\top f(z)) \quad (6)$$

For a fixed value of $\lambda$ the optimal $q(z) \propto q'(z) \exp(-\lambda^\top f(z))$. By updating $q(\theta)$ and $q(z)$ as in (2) and (3) we are effectively maximizing the lower bound $\mathcal{F}$.

[2] Constraints of the form $E_q[f(z)] \geq b$ are easily imposed by negating $f(z)$ and $b$.

### 4.1 Variational Updates

We now derive the specific variational updates for our dependency induction model. First we assume the following mean-field factorization of our variational distribution:

$$\begin{aligned} & q(\beta, \theta, \pi, \phi, z) \\ & = q(z) \cdot \prod_{s'} q(\beta_{s'}) \cdot \prod_{z'=1}^{T} q(\phi_{s'z'}) \cdot \\ & \quad \prod_{c} q(\theta_{s'z'c}) \cdot \prod_{s} q(\pi_{ss'z'c}), \end{aligned} \tag{7}$$

where $s'$ varies over the set of unique symbols in the observed tags, $z'$ denotes subsymbols for each symbol, $c$ varies over context values comprising a pair of direction (left or right) and valence (first, second, or third or higher) values, and $s$ corresponds to child symbols.

We restrict $q(\theta_{s'z'c})$ and $q(\phi_{s'z'})$ to be Dirichlet distributions and $q(z)$ to be multinomial. As with prior work (Liang et al., 2009b), we assume a degenerate $q(\beta) \equiv \delta_{\beta^*}(\beta)$ for tractability reasons, i.e., all mass is concentrated on some single $\beta^*$. We also assume that the top level stick-breaking distribution is truncated at $T$, i.e., $q(\beta)$ assigns zero probability to integers greater than $T$. Because of the truncation of $\beta$, we can approximate $q(\pi_{ss'z'c})$ with an asymmetric finite dimensional Dirichlet.

The factors are updated one at a time holding all other factors fixed. The variational update for $q(\pi)$ is given by:

$$q(\pi_{ss'z'c}) = \mathrm{Dir}\left(\pi_{ss'z'c}; \alpha\beta + E_{q(z)}[C_{ss'z'c}(z)]\right),$$

where term $E_{q(z)}[C_{ss'z'c}(z)]$ is the expected count w.r.t. $q(z)$ of child symbol $s$ and subsymbol $z$ in context $c$ when generated by parent symbol $s'$ and subsymbol $z'$.

Similarly, the updates for $q(\theta)$ and $q(\phi)$ are given by:

$$\begin{aligned} q(\theta_{s'z'c}) &= \mathrm{Dir}\left(\theta_{s'z'c}; \theta_0 + E_{q(z)}[C_{s'z'c}(s)]\right), \\ q(\phi_{s'z'}) &= \mathrm{Dir}\left(\phi_{s'z'}; \phi_0 + E_{q(z)}[C_{s'z'}(x)]\right), \end{aligned}$$

where $C_{s'z'c}(s)$ is the count of child symbol $s$ being generated by the parent symbol $s'$ and subsymbol $z'$ in context $c$ and $C_{s'z'x}$ is the count of word $x$ being generated by symbol $s'$ and subsymbol $z'$.

The only factor affected by the expectation constraints is $q(z)$. Recall from the previous section that the update for $q(z)$ is performed via gradient search on the dual of a constrained minimization problem of the form:

$$q(z) = \operatorname*{argmin}_{q(z)} \mathrm{KL}(q(z) \parallel q'(z)).$$

Thus we first compute the update for $q'(z)$:

$$\begin{aligned} q'(z) \propto \prod_{n=1}^{N} \prod_{j=1}^{len(n)} & (\exp E_{q(\phi)}[\log \phi_{s_{nj} z_{nj}}(x_{nj})] \\ & \times \exp E_{q(\theta)}[\log \theta_{s_{h(nj)} z_{h(nj)} c_{nj}}(s_{nj})] \\ & \times \exp E_{q(\pi)}[\log \pi_{s_{nj} s_{h(nj)} z_{h(nj)} c_{nj}}(z_{nj})]), \end{aligned}$$

where $N$ is the total number of sentences, $len(n)$ is the length of sentence $n$, and index $h(nj)$ refers to the head of the $j$th node of sentence $n$. Given this $q'(z)$ a gradient search is performed using (6) to find the optimal $\lambda$ and thus the primal solution for updating $q(z)$.

Finally, we update the degenerate factor $q(\beta_s)$ with the projected gradient search algorithm used by Liang et al. (2009b).

## 5 Linguistic Constraints

**Universal Dependency Rules** We compile a set of 13 universal dependency rules consistent with various linguistic accounts (Carnie, 2002; Newmeyer, 2005), shown in Table 1. These rules are defined over coarse part-of-speech tags: Noun, Verb, Adjective, Adverb, Pronoun, Article, Auxiliary, Preposition, Numeral and Conjunction. Each rule specifies a part-of-speech for the head and argument but does not provide ordering information.

We require that a minimum proportion of the posterior dependencies be instances of these rules in expectation. In contrast to prior work on rule-driven dependency induction (Druck et al., 2009), where each rule has a separately specified expectation, we only set a single minimum expectation for the proportion of all dependencies that must match one of the rules. This setup is more relevant for learning with universals since individual rule frequencies vary greatly between languages.

1. Identify non-recursive NPs:
   - All nouns, pronouns and possessive marker are part of an NP.
   - All adjectives, conjunctions and determiners immediately preceding an NP are part of the NP.
2. The first verb or modal in the sentence is the headword.
3. All words in an NP are headed by the last word in the NP.
4. The last word in an NP is headed by the word immediately before the NP if it is a preposition, otherwise it is headed by the headword of the sentence if the NP is before the headword, else it is headed by the word preceding the NP.
5. For the first word set its head to be the headword of the sentence. For each other word set its headword to be the previous word.

Table 3: English-specific dependency rules.

**English-specific Dependency Rules** For English, we also consider a small set of hand-crafted dependency rules designed by Michael Collins[3] for deterministic parsing, shown in Table 3. Unlike the universals from Table 1, these rules alone are enough to construct a full dependency tree. Thus they allow us to judge whether the model is able to improve upon a human-engineered deterministic parser. Moreover, with this dataset we can assess the additional benefit of using rules tailored to an individual language as opposed to universal rules.

## 6 Experimental Setup

**Datasets and Evaluation** We test the effectiveness of our grammar induction approach on English, Danish, Portuguese, Slovene, Spanish, and Swedish. For English we use the Penn Treebank (Marcus et al., 1993), transformed from CFG parses into dependencies with the Collins head finding rules (Collins, 1999); for the other languages we use data from the 2006 CoNLL-X Shared Task (Buchholz and Marsi, 2006). Each dataset provides manually annotated part-of-speech tags that are used for both training and testing. For comparison purposes with previous work, we limit the cross-lingual experiments to sentences of length 10 or less (not counting punctuation). For English, we also explore sentences of length up to 20.

The final output metric is directed dependency accuracy. This is computed based on the Viterbi parses produced using the final unnormalized variational distribution $q(z)$ over dependency structures.

**Hyperparameters and Training Regimes** Unless otherwise stated, in experiments with rule-based constraints the expected proportion of dependencies that must satisfy those constraints is set to 0.8. This threshold value was chosen based on minimal tuning on a single language and ruleset (English with universal rules) and carried over to each other experimental condition. A more detailed discussion of the threshold's empirical impact is presented in Section 7.1.

Variational approximations to the HDP are truncated at 10. All hyperparameter values are fixed to 1 except $\alpha$ which is fixed to 10.

We also conduct a set of *No-Split* experiments to evaluate the importance of syntactic refinement; in these experiments each coarse symbol corresponds to only one refined symbol. This is easily effected during inference by setting the HDP variational approximation truncation level to one.

For each experiment we run 50 iterations of variational updates; for each iteration we perform five steps of gradient search to compute the update for the variational distribution $q(z)$ over dependency structures.

## 7 Results

In the following section we present our primary cross-lingual results using universal rules (Section 7.1) before performing a more in-depth analysis of model properties such as sensitivity to ruleset selection and inference stability (Section 7.2).

[3] Personal communication.

| | DMV | PGI | No-Split | HDP-DEP |
|---|---|---|---|---|
| English | 47.1 | 62.3 | 71.5 | **71.9** (0.3) |
| Danish | 33.5 | 41.6 | 48.8 | **51.9** (1.6) |
| Portuguese | 38.5 | 63.0 | 54.0 | **71.5** (0.5) |
| Slovene | 38.5 | 48.4 | 50.6 | **50.9** (5.5) |
| Spanish | 28.0 | 58.4 | 64.8 | **67.2** (0.4) |
| Swedish | 45.3 | 58.3 | **63.3** | 62.1 (0.5) |

Table 4: Directed dependency accuracy using our model with universal dependency rules (No-Split and HDP-DEP), compared to DMV (Klein and Manning, 2004) and PGI (Berg-Kirkpatrick and Klein, 2010). The DMV results are taken from Berg-Kirkpatrick and Klein (2010). Bold numbers indicate the best result for each language. For the full model, the standard deviation in performance over five runs is indicated in parentheses.

| English | | | |
|---|---|---|---|
| Rule Excluded | Acc | Loss | Gold Freq |
| Preposition → Noun | 61.0 | 10.9 | 5.1 |
| Verb → Noun | 61.4 | 10.5 | 14.8 |
| Noun → Noun | 64.4 | 7.5 | 10.7 |
| Noun → Article | 64.7 | 7.2 | 8.5 |
| **Spanish** | | | |
| Rule Excluded | Acc | Loss | Gold Freq |
| Preposition → Noun | 53.4 | 13.8 | 8.2 |
| Verb → Noun | 61.9 | 5.4 | 12.9 |
| Noun → Noun | 62.6 | 4.7 | 2.0 |
| Root → Verb | 65.4 | 1.8 | 12.3 |

Table 5: Ablation experiment results for universal dependency rules on English and Spanish. For each rule, we evaluate the model using the ruleset excluding that rule, and list the most significant rules for each language. The second last column is the absolute loss in performance compared to the setting where all rules are available. The last column shows the percentage of the gold dependencies that satisfy the rule.

### 7.1 Main Cross-Lingual Results

Table 4 shows the performance of both our full model (HDP-DEP) and its No-Split version using universal dependency rules across six languages. We also provide the performance of two baselines — the dependency model with valence (DMV) (Klein and Manning, 2004) and the phylogenetic grammar induction (PGI) model (Berg-Kirkpatrick and Klein, 2010).

HDP-DEP outperforms both DMV and PGI across all six languages. Against DMV we achieve an average absolute improvement of 24.1%. This improvement is expected given that DMV does not have access to the additional information provided through the universal rules. PGI is more relevant as a point of comparison, since it is able to leverage multilingual data to learn information similar to what we have declaratively specified using universal rules. Specifically, PGI reduces induction ambiguity by connecting language-specific parameters via phylogenetic priors. We find, however, that we outperform PGI by an average margin of 7.2%, demonstrating the benefits of explicit rule specification.

An additional point of comparison is the lexicalized unsupervised parser of Headden III et al. (2009), which yields the current state-of-the-art unsupervised accuracy on English at 68.8%. Our method also outperforms this approach, without employing lexicalization and sophisticated smoothing as they do. This result suggests that combining the complementary strengths of their approach and ours can yield further performance improvements.

Table 4 also shows the *No-Split* results where syntactic categories are not refined. We find that such refinement usually proves to be beneficial, yielding an average performance gain of 3.7%. However, we note that the impact of incorporating splitting varies significantly across languages. Further understanding of this connection is an area of future research.

Finally, we note that our model exhibits low variance for most languages. This result attests to how the expectation constraints consistently guide inference toward high-accuracy areas of the search space.

**Ablation Analysis** Our next experiment seeks to understand the relative importance of the various universal rules from Table 1. We study how accuracy is affected when each of the rules is removed one at a time for English and Spanish. Table 5 lists the rules with the greatest impact on performance when removed. We note the high overlap between the most significant rules for English and Spanish.

We also observe that the relationship between a rule's frequency and its importance for high accuracy is not straightforward. For example, the "Preposition → Noun" rule, whose removal degrades accuracy the most for both English and Span-

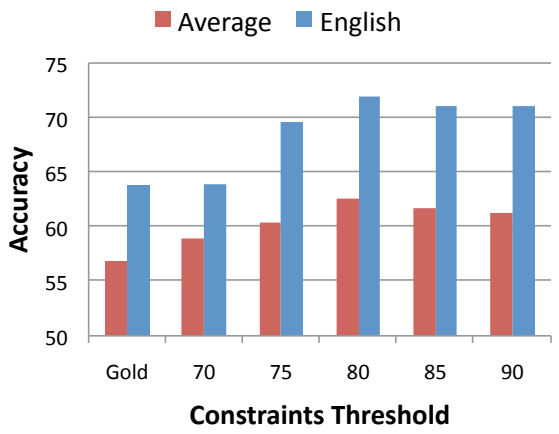



Figure 2: Accuracy of our model with different threshold settings, on English only and averaged over all languages. "Gold" refers to the setting where each language's threshold is set independently to the proportion of gold dependencies satisfying the rules — for English this proportion is 70%, while the average proportion across languages is 63%.

| | | Length | |
|---|---|---|---|
| | | $\leq 10$ | $\leq 20$ |
| Universal Dependency Rules | | | |
| **1** | HDP-DEP | 71.9 | 50.4 |
| No Rules (Random Init) | | | |
| **2** | HDP-DEP | 24.9 | 24.4 |
| **3** | Headden III et al. (2009) | 68.8 | - |
| English-Specific Parsing Rules | | | |
| **4** | Deterministic (rules only) | 70.0 | 62.6 |
| **5** | HDP-DEP | **73.8** | **66.1** |
| Druck et al. (2009) Rules | | | |
| **6** | Druck et al. (2009) | 61.3 | - |
| **7** | HDP-DEP | 64.9 | 42.2 |

Table 6: Directed accuracy of our model (HDP-DEP) on sentences of length 10 or less and 20 or less from WSJ with different rulesets and with no rules, along with various baselines from the literature. Entries in this table are numbered for ease of reference in the text.

ish, is not the most frequent rule in either language. This result suggests that some rules are harder to learn than others regardless of their frequency, so their presence in the specified ruleset yields stronger performance gains.

**Varying the Constraint Threshold** In our main experiments we require that at least 80% of the expected dependencies satisfy the rule constraints. We arrived at this threshold by tuning on the basis of English only. As shown in Figure 2, for English a broad band of threshold values from 75% to 90% yields results within 2.5% of each other, with a slight peak at 80%.

To further study the sensitivity of our method to how the threshold is set, we perform *post hoc* experiments with other threshold values on each of the other languages. As Figure 2 also shows, on average a value of 80% is optimal across languages, though again accuracy is stable within 2.5% between thresholds of 75% to 90%. These results demonstrate that a single threshold is broadly applicable across languages.

Interestingly, setting the threshold value independently for each language to its "true" proportion based on the gold dependencies (denoted as the "Gold" case in Figure 2) does not achieve optimal performance. Thus, knowledge of the true language-specific rule proportions is not necessary for high accuracy.

### 7.2 Analysis of Model Properties

We perform a set of additional experiments on English to gain further insight into HDP-DEP's behavior. Our choice of language is motivated by the fact that a wide range of prior parsing algorithms were developed for and tested exclusively on English. The experiments below demonstrate that 1) universal rules alone are powerful, but language- and dataset-tailored rules can further improve performance; 2) our model learns jointly from the rules and data, outperforming a rules-only deterministic parser; 3) the way we incorporate posterior constraints outperforms the generalized expectation constraint framework; and 4) our model exhibits low variance when seeded with different initializations. These results are summarized in Table 6 and discussed in detail below; line numbers refer to entries in Table 6. Each run of HDP-DEP below is with syntactic refinement enabled.

**Impact of Rules Selection** We compare the performance of HDP-DEP using the universal rules versus a set of rules designed for deterministically parsing the Penn Treebank (see Section 5 for details).

As lines 1 and 5 of Table 6 show, language-specific rules yield better performance. For sentences of length 10 or less, the difference between the two rulesets is a relatively small 1.9%; for longer sentences, however, the difference is a substantially larger 15.7%. This is likely because longer sentences tend to be more complex and thus exhibit more language-idiosyncratic dependencies. Such dependencies can be better captured by the refined language-specific rules.

We also test model performance when no linguistic rules are available, i.e., performing unconstrained variational inference. The model performs substantially worse (line 2), confirming that syntactic category refinement in a fully unsupervised setup is challenging.

**Learning Beyond Provided Rules** Since HDP-DEP is provided with linguistic rules, a legitimate question is whether it improves upon what the rules encode, especially when the rules are complete and language-specific. We can answer this question by comparing the performance of our model seeded with the English-specific rules against a deterministic parser that implements the same rules. Lines 4 and 5 of Table 6 demonstrate that the model outperforms a rules-only deterministic parser by 3.8% for sentences of length 10 or less and by 3.5% for sentences of length 20 or less.

**Comparison with Alternative Semi-supervised Parser** The dependency parser based on the generalized expectation criteria (Druck et al., 2009) is the closest to our reported work in terms of technique. To compare the two, we run HDP-DEP using the 20 rules given by Druck et al. (2009). Our model achieves an accuracy of 64.9% (line 7) compared to 61.3% (line 6) reported in their work. Note that we do not rely on rule-specific expectation information as they do, instead requiring only a single expectation constraint parameter.[4]

**Model Stability** It is commonly acknowledged in the literature that unsupervised grammar induction methods exhibit sensitivity to initialization. As in the previous section, we find that the presence of linguistic rules greatly reduces this sensitivity: for HDP-DEP, the standard deviation over five randomly initialized runs with the English-specific rules is 1.5%, compared to 4.5% for the parser developed by Headden III et al. (2009) and 8.0% for DMV (Klein and Manning, 2004).

[4] As explained in Section 5, having a single expectation parameter is motivated by our focus on parsing with universal rules.

## 8 Conclusions

In this paper we demonstrated that syntactic universals encoded as declarative constraints improve grammar induction. We formulated a generative model for dependency structure that models syntactic category refinement and biases inference to cohere with the provided constraints. Our experiments showed that encoding a compact, well-accepted set of language-independent constraints significantly improves accuracy on multiple languages compared to the current state-of-the-art in unsupervised parsing.

While our present work has yielded substantial gains over previous unsupervised methods, a large gap still remains between our method and fully supervised techniques. In future work we intend to study ways to bridge this gap by 1) incorporating more sophisticated linguistically-driven grammar rulesets to guide induction, 2) lexicalizing the model, and 3) combining our constraint-based approach with richer unsupervised models (e.g., Headden III et al. (2009)) to benefit from their complementary strengths.

## Acknowledgments

The authors acknowledge the support of the NSF (CAREER grant IIS-0448168, grant IIS-0904684, and a Graduate Research Fellowship). We are especially grateful to Michael Collins for inspiring us toward this line of inquiry and providing deterministic rules for English parsing. Thanks to Taylor Berg-Kirkpatrick, Sabine Iatridou, Ramesh Sridharan, and members of the MIT NLP group for their suggestions and comments. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors, and do not necessarily reflect the views of the funding organizations.

## References


Mark C. Baker. 2001. *The Atoms of Language: The Mind's Hidden Rules of Grammar*. Basic Books.

Emily M. Bender. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32.

Taylor Berg-Kirkpatrick and Dan Klein. 2010. Phylogenetic grammar induction. In *Proceedings of ACL*, pages 1288–1297.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of CoNLL*, pages 149–164.

David Burkett and Dan Klein. 2008. Two languages are better than one (for syntactic parsing). In *Proceedings of EMNLP*, pages 877–886.

Andrew Carnie. 2002. *Syntax: A Generative Introduction (Introducing Linguistics)*. Blackwell Publishing.

Ming-Wei Chang, Lev Ratinov, and Dan Roth. 2007. Guiding semi-supervision with constraint-driven learning. In *Proceedings of ACL*, pages 280–287.

Shay B. Cohen and Noah A. Smith. 2009a. Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In *Proceedings of NAACL/HLT*, pages 74–82.

Shay B. Cohen and Noah A. Smith. 2009b. Variational inference for grammar induction with prior knowledge. In *Proceedings of ACL/IJCNLP 2009 Conference Short Papers*, pages 1–4.

Michael Collins. 1999. *Head-driven statistical models for natural language parsing*. Ph.D. thesis, University of Pennsylvania.

Hal Daumé III and Lyle Campbell. 2007. A bayesian model for discovering typological implications. In *Proceedings of ACL*, pages 65–72.

Gregory Druck, Gideon Mann, and Andrew McCallum. 2009. Semi-supervised learning of dependency parsers using generalized expectation criteria. In *Proceedings of ACL/IJCNLP*, pages 360–368.

Thomas S. Ferguson. 1973. A bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230.

Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. 2007. The infinite tree. In *Proceedings of ACL*, pages 272–279.

Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. 2009. Dependency grammar induction via bitext projection constraints. In *Proceedings of ACL/IJCNLP*, pages 369–377.

Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049.

João Graça, Kuzman Ganchev, Ben Taskar, and Fernando Pereira. 2009. Posterior vs. parameter sparsity in latent variable models. In *Advances in NIPS*, pages 664–672.

João Graça, Kuzman Ganchev, and Ben Taskar. 2007. Expectation maximization and posterior constraints. In *Advances in NIPS*, pages 569–576.

Aria Haghighi and Dan Klein. 2006. Prototype-driven grammar induction. In *Proceedings of ACL*, pages 881–888.

William P. Headden III, Mark Johnson, and David McClosky. 2009. Improving unsupervised dependency parsing with richer contexts and smoothing. In *Proceedings of NAACL/HLT*, pages 101–109.

Dan Klein and Christopher Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of ACL*, pages 478–485.

Jonas Kuhn. 2004. Experiments in parallel-text based grammar induction. In *Proceedings of ACL*, pages 470–477.

Percy Liang, Slav Petrov, Michael Jordan, and Dan Klein. 2007. The infinite PCFG using hierarchical Dirichlet processes. In *Proceedings of EMNLP/CoNLL*, pages 688–697.

Percy Liang, Michael I. Jordan, and Dan Klein. 2009a. Learning from measurements in exponential families. In *Proceedings of ICML*, pages 641–648.

Percy Liang, Michael I. Jordan, and Dan Klein. 2009b. Probabilistic grammars and hierarchical Dirichlet processes. *The Handbook of Applied Bayesian Analysis.*

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.

Frederick J. Newmeyer. 2005. *Possible and Probable Languages: A Generative Perspective on Linguistic Typology*. Oxford University Press.

Slav Petrov and Dan Klein. 2007. Learning and inference for hierarchically split PCFGs. In *Proceeding of AAAI*, pages 1663–1666.

Benjamin Snyder, Tahira Naseem, and Regina Barzilay. 2009. Unsupervised multilingual grammar induction. In *Proceedings of ACL/IJCNLP*, pages 73–81.

Lydia White. 2003. *Second Language Acquisition and Universal Grammar*. Cambridge University Press.