Learning Common Grammar from Multilingual Corpus

Tomoharu Iwata **Daichi Mochihashi** **Hiroshi Sawada**
NTT Communication Science Laboratories
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan
{iwata,daichi,sawada}@cslab.kecl.ntt.co.jp

Abstract

We propose a corpus-based probabilistic framework to extract hidden common syntax across languages from non-parallel multilingual corpora in an unsupervised fashion. For this purpose, we assume a generative model for multilingual corpora, where each sentence is generated from a language dependent probabilistic context-free grammar (PCFG), and these PCFGs are generated from a prior grammar that is common across languages. We also develop a variational method for efficient inference. Experiments on a non-parallel multilingual corpus of eleven languages demonstrate the feasibility of the proposed method.

1 Introduction

Languages share certain common properties (Pinker, 1994). For example, the word order in most European languages is subject-verb-object (SVO), and some words with similar forms are used with similar meanings in different languages. The reasons for these common properties can be attributed to: 1) a common ancestor language, 2) borrowing from nearby languages, and 3) the innate abilities of humans (Chomsky, 1965).

We assume hidden commonalities in syntax across languages, and try to extract a common grammar from non-parallel multilingual corpora. For this purpose, we propose a generative model for multilingual grammars that is learned in an unsupervised fashion. There are some computational models for capturing commonalities at the phoneme and word level (Oakes, 2000; Bouchard-Côté et al., 2008), but, as far as we know, no attempt has been made to extract commonalities in syntax level from non-parallel and non-annotated multilingual corpora.

In our scenario, we use probabilistic context-free grammars (PCFGs) as our monolingual grammar model. We assume that a PCFG for each language is generated from a general model that are common across languages, and each sentence in multilingual corpora is generated from the language dependent PCFG. The inference of the general model as well as the multilingual PCFGs can be performed by using a variational method for efficiency. Our approach is based on a Bayesian multitask learning framework (Yu et al., 2005; Daumé III, 2009). Hierarchical Bayesian modeling provides a natural way of obtaining a joint regularization for individual models by assuming that the model parameters are drawn from a common prior distribution (Yu et al., 2005).

2 Related work

The unsupervised grammar induction task has been extensively studied (Carroll and Charniak, 1992; Stolcke and Omohundro, 1994; Klein and Manning, 2002; Klein and Manning, 2004; Liang et al., 2007). Recently, models have been proposed that outperform PCFG in the grammar induction task (Klein and Manning, 2002; Klein and Manning, 2004). We used PCFG as a first step for capturing commonalities in syntax across languages because of its simplicity. The proposed framework can be used for probabilistic grammar models other than PCFG.

Grammar induction using bilingual parallel corpora has been studied mainly in machine translation research (Wu, 1997; Melamed, 2003; Eisner, 2003; Chiang, 2005; Blunsom et al., 2009; Snyder et al., 2009). These methods require sentence-aligned parallel data, which can be costly to obtain and difficult to scale to many languages. On the other hand, our model does not require sentences to be aligned. Moreover, since the complexity of our model increases linearly with the number of languages, our model is easily applicable to cor-

Proceedings of the ACL 2010 Conference Short Papers, pages 184–188,
Uppsala, Sweden, 11-16 July 2010. ©2010 Association for Computational Linguistics

pora of more than two languages, as we will show in the experiments. To our knowledge, the only grammar induction work on non-parallel corpora is (Cohen and Smith, 2009), but their method does not model a common grammar, and requires prior information such as part-of-speech tags. In contrast, our method does not require any such prior information.

3 Proposed Method

3.1 Model

Let $\boldsymbol{X} = \{\boldsymbol{X}_l\}_{l \in \boldsymbol{L}}$ be a non-parallel and non-annotated multilingual corpus, where $\boldsymbol{X}_l$ is a set of sentences in language l, and $\boldsymbol{L}$ is a set of languages. The task is to learn multilingual PCFGs $\boldsymbol{G} = \{\boldsymbol{G}_l\}_{l \in \boldsymbol{L}}$ and a common grammar that generates these PCFGs. Here, $\boldsymbol{G}_l = (\boldsymbol{K}, \boldsymbol{W}_l, \boldsymbol{\Phi}_l)$ represents a PCFG of language l, where $\boldsymbol{K}$ is a set of nonterminals, $\boldsymbol{W}_l$ is a set of terminals, and $\boldsymbol{\Phi}_l$ is a set of rule probabilities. Note that a set of nonterminals $\boldsymbol{K}$ is shared among languages, but a set of terminals $\boldsymbol{W}_l$ and rule probabilities $\boldsymbol{\Phi}_l$ are specific to the language. For simplicity, we consider Chomsky normal form grammars, which have two types of rules: emissions rewrite a nonterminal as a terminal $A \rightarrow w$, and binary productions rewrite a nonterminal as two nonterminals $A \rightarrow BC$, where $A, B, C \in \boldsymbol{K}$ and $w \in \boldsymbol{W}_l$.

The rule probabilities for each nonterminal A of PCFG G_l in language l consist of: 1) $\boldsymbol{\theta}_{Al} = \{\theta_{lAt}\}_{t \in \{0,1\}}$, where θ_{lA0} and θ_{lA1} represent probabilities of choosing the emission rule and the binary production rule, respectively, 2) $\boldsymbol{\phi}_{lA} = \{\phi_{lABC}\}_{B,C \in \boldsymbol{K}}$, where ϕ_{lABC} represents the probability of nonterminal production $A \rightarrow BC$, and 3) $\boldsymbol{\psi}_{lA} = \{\psi_{lAw}\}_{w \in \boldsymbol{W}_l}$, where ψ_{lAw} represents the probability of terminal emission $A \rightarrow w$. Note that $\theta_{lA0} + \theta_{lA1} = 1$, $\theta_{lAt} \geq 0$, $\sum_{B,C} \phi_{lABC} = 1$, $\phi_{lABC} \geq 0$, $\sum_w \psi_{lAw} = 1$, and $\psi_{lAw} \geq 0$. In the proposed model, multinomial parameters $\boldsymbol{\theta}_{lA}$ and $\boldsymbol{\phi}_{lA}$ are generated from Dirichlet distributions that are common across languages: $\theta_{lA} \sim \mathrm{Dir}(\boldsymbol{\alpha}^\theta_A)$ and $\boldsymbol{\phi}_{lA} \sim \mathrm{Dir}(\boldsymbol{\alpha}^\phi_A)$, since we assume that languages share a common syntax structure. $\boldsymbol{\alpha}^\theta_A$ and $\boldsymbol{\alpha}^\phi_A$ represent the parameters of a common grammar. We use the Dirichlet prior because it is the conjugate prior for the multinomial distribution. In summary, the proposed model assumes the following generative process for a multilingual corpus,

1. For each nonterminal $A \in \boldsymbol{K}$:

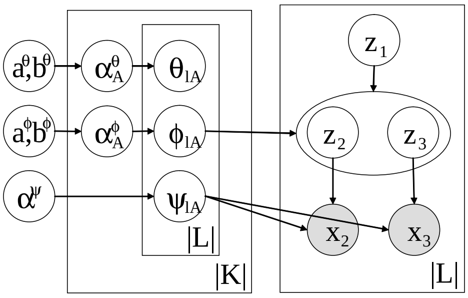

Figure 1: Graphical model.

   (a) For each rule type $t \in \{0, 1\}$:
      i. Draw common rule type parameters $\alpha^\theta_{At} \sim \mathrm{Gam}(a^\theta, b^\theta)$
   (b) For each nonterminal pair (B, C):
      i. Draw common production parameters $\alpha^\phi_{ABC} \sim \mathrm{Gam}(a^\phi, b^\phi)$

2. For each language $l \in \boldsymbol{L}$:

   (a) For each nonterminal $A \in \boldsymbol{K}$:
      i. Draw rule type parameters $\theta_{lA} \sim \mathrm{Dir}(\boldsymbol{\alpha}^\theta_A)$
      ii. Draw binary production parameters $\boldsymbol{\phi}_{lA} \sim \mathrm{Dir}(\boldsymbol{\alpha}^\phi_A)$
      iii. Draw emission parameters $\boldsymbol{\psi}_{lA} \sim \mathrm{Dir}(\alpha^\psi)$
   (b) For each node i in the parse tree:
      i. Choose rule type $t_{li} \sim \mathrm{Mult}(\boldsymbol{\theta}_{lz_i})$
      ii. If $t_{li} = 0$:
         A. Emit terminal $x_{li} \sim \mathrm{Mult}(\boldsymbol{\psi}_{lz_i})$
      iii. Otherwise:
         A. Generate children nonterminals $(z_{lL(i)}, z_{lR(i)}) \sim \mathrm{Mult}(\boldsymbol{\phi}_{lz_i})$,

where $L(i)$ and $R(i)$ represent the left and right children of node i. Figure 1 shows a graphical model representation of the proposed model, where the shaded and unshaded nodes indicate observed and latent variables, respectively.

3.2 Inference

The inference of the proposed model can be efficiently computed using a variational Bayesian method. We extend the variational method to the monolingual PCFG learning of Kurihara and Sato (2004) for multilingual corpora. The goal is to estimate posterior $p(\boldsymbol{Z}, \boldsymbol{\Phi}, \boldsymbol{\alpha}|\boldsymbol{X})$, where $\boldsymbol{Z}$ is a set of parse trees, $\boldsymbol{\Phi} = \{\boldsymbol{\Phi}_l\}_{l \in \boldsymbol{L}}$ is a set of language dependent parameters, $\boldsymbol{\Phi}_l = \{\boldsymbol{\theta}_{lA}, \boldsymbol{\phi}_{lA}, \boldsymbol{\psi}_{lA}\}_{A \in \boldsymbol{K}}$, and $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}^\theta_A, \boldsymbol{\alpha}^\phi_A\}_{A \in \boldsymbol{K}}$ is a set of common parameters. In the variational method, posterior $p(\boldsymbol{Z}, \boldsymbol{\Phi}, \boldsymbol{\alpha}|\boldsymbol{X})$ is approximated by a tractable variational distribution $q(\boldsymbol{Z}, \boldsymbol{\Phi}, \boldsymbol{\alpha})$.

We use the following variational distribution,

$$\begin{aligned} q(\boldsymbol{Z},\boldsymbol{\Phi},\boldsymbol{\alpha}) &= \prod_{A} q(\boldsymbol{\alpha}^{\theta}_{A}) q(\boldsymbol{\alpha}^{\phi}_{A}) \prod_{l,d} q(\boldsymbol{z}_{ld}) \\ &\times \prod_{l,A} q(\boldsymbol{\theta}_{lA}) q(\boldsymbol{\phi}_{lA}) q(\boldsymbol{\psi}_{lA}), \quad (1) \end{aligned}$$

where we assume that hyperparameters $q(\boldsymbol{\alpha}^{\theta}_{A})$ and $q(\boldsymbol{\alpha}^{\phi}_{A})$ are degenerated, or $q(\boldsymbol{\alpha}) = \delta_{\boldsymbol{\alpha}^{*}}(\boldsymbol{\alpha})$, and infer them by point estimation instead of distribution estimation. We find an approximate posterior distribution that minimizes the Kullback-Leibler divergence from the true posterior. The variational distribution of the parse tree of the dth sentence in language l is obtained as follows,

$$\begin{aligned} q(\boldsymbol{z}_{ld}) &\propto \prod_{A\rightarrow BC} \left(\pi^{\theta}_{lA1}\pi^{\phi}_{lABC}\right)^{C(A\rightarrow BC;\boldsymbol{z}_{ld},l,d)} \\ &\times \prod_{A\rightarrow w} \left(\pi^{\theta}_{lA0}\pi^{\psi}_{lAw}\right)^{C(A\rightarrow w;\boldsymbol{z}_{ld},l,d)}, \quad (2) \end{aligned}$$

where $C(r;\boldsymbol{z},l,d)$ is the count of rule r that occurs in the dth sentence of language l with parse tree $\boldsymbol{z}$. The multinomial weights are calculated as follows,

$$\pi^{\theta}_{lAt} = \exp\bigl(\mathbb{E}_{q(\boldsymbol{\theta}_{lA})}\bigl[\log\theta_{lAt}\bigr]\bigr), \quad (3)$$

$$\pi^{\phi}_{lABC} = \exp\bigl(\mathbb{E}_{q(\boldsymbol{\phi}_{lA})}\bigl[\log\phi_{lABC}\bigr]\bigr), \quad (4)$$

$$\pi^{\psi}_{lAw} = \exp\bigl(\mathbb{E}_{q(\boldsymbol{\psi}_{lA})}\bigl[\log\psi_{lAw}\bigr]\bigr). \quad (5)$$

The variational Dirichlet parameters for $q(\boldsymbol{\theta}_{lA}) = \mathrm{Dir}(\boldsymbol{\gamma}^{\theta}_{lA})$, $q(\boldsymbol{\phi}_{lA}) = \mathrm{Dir}(\boldsymbol{\gamma}^{\phi}_{lA})$, and $q(\boldsymbol{\psi}_{lA}) = \mathrm{Dir}(\boldsymbol{\gamma}^{\psi}_{lA})$, are obtained as follows,

$$\gamma^{\theta}_{lAt} = \alpha^{\theta}_{At} + \sum_{d,\boldsymbol{z}_{ld}} q(\boldsymbol{z}_{ld}) C(A,t;\boldsymbol{z}_{ld},l,d), \quad (6)$$

$$\gamma^{\phi}_{lABC} = \alpha^{\phi}_{ABC} + \sum_{d,\boldsymbol{z}_{ld}} q(\boldsymbol{z}_{ld}) C(A\rightarrow BC;\boldsymbol{z}_{ld},l,d), \quad (7)$$

$$\gamma^{\psi}_{lAw} = \alpha^{\psi} + \sum_{d,\boldsymbol{z}_{ld}} q(\boldsymbol{z}_{ld}) C(A\rightarrow w;\boldsymbol{z}_{ld},l,d), \quad (8)$$

where $C(A,t;\boldsymbol{z},l,d)$ is the count of rule type t that is selected in nonterminal A in the dth sentence of language l with parse tree $\boldsymbol{z}$.

The common rule type parameter α^{θ}_{At} that minimizes the KL divergence between the true posterior and the approximate posterior can be obtained by using the fixed-point iteration method described in (Minka, 2000). The update rule is as follows,

$$\alpha^{\theta(\mathrm{new})}_{At} \leftarrow \frac{a^{\theta}-1+\alpha^{\theta}_{At}L\bigl(\Psi(\sum_{t'}\alpha^{\theta}_{At'})-\Psi(\alpha^{\theta}_{At})\bigr)}{b^{\theta}+\sum_{l}\bigl(\Psi(\sum_{t'}\gamma^{\theta}_{lAt'})-\Psi(\gamma^{\theta}_{lAt})\bigr)}, \quad (9)$$

where L is the number of languages, and $\Psi(x) = \frac{\partial\log\Gamma(x)}{\partial x}$ is the digamma function. Similarly, the common production parameter α^{ϕ}_{ABC} can be updated as follows,

$$\alpha^{\phi(\mathrm{new})}_{ABC} \leftarrow \frac{a^{\phi}-1+\alpha^{\phi}_{ABC}LJ_{ABC}}{b^{\phi}+\sum_{l}J'_{lABC}}, \quad (10)$$

where $J_{ABC} = \Psi(\sum_{B',C'}\alpha^{\phi}_{AB'C'}) - \Psi(\alpha^{\phi}_{ABC})$, and $J'_{lABC} = \Psi(\sum_{B',C'}\gamma^{\phi}_{lAB'C'}) - \Psi(\gamma^{\phi}_{lABC})$.

Since factored variational distributions depend on each other, an optimal approximated posterior can be obtained by updating parameters by (2) - (10) alternatively until convergence. The updating of language dependent distributions by (2) - (8) is also described in (Kurihara and Sato, 2004; Liang et al., 2007) while the updating of common grammar parameters by (9) and (10) is new. The inference can be carried out efficiently using the inside-outside algorithm based on dynamic programming (Lari and Young, 1990).

After the inference, the probability of a common grammar rule $A \rightarrow BC$ is calculated by $\hat{\phi}_{A\rightarrow BC} = \hat{\theta}_{1}\hat{\phi}_{ABC}$, where $\hat{\theta}_{1} = \alpha^{\theta}_{1}/(\alpha^{\theta}_{0}+\alpha^{\theta}_{1})$ and $\hat{\phi}_{ABC} = \alpha^{\phi}_{ABC}/\sum_{B',C'}\alpha^{\phi}_{AB'C'}$ represent the mean values of θ_{l0} and ϕ_{lABC}, respectively.

4 Experimental results

We evaluated our method by employing the EuroParl corpus (Koehn, 2005). The corpus consists of the proceedings of the European Parliament in eleven western European languages: Danish (da), German (de), Greek (el), English (en), Spanish (es), Finnish (fi), French (fr), Italian (it), Dutch (nl), Portuguese (pt), and Swedish (sv), and it contains roughly 1,500,000 sentences in each language. We set the number of nonterminals at $|\boldsymbol{K}| = 20$, and omitted sentences with more than ten words for tractability. We randomly sampled 100,000 sentences for each language, and analyzed them using our method. It should be noted that our random samples are not sentence-aligned.

Figure 2 shows the most probable terminals of emission for each language and nonterminal with a high probability of selecting the emission rule.

2: verb and auxiliary verb (V)

```
[da] det jeg vi der de derfor dette forhandlingen hvad
[de] ist sind haben wird hat müssen möchte meine werden kann
[el] , είναι για να πρέπει δεν και αυτό με ότι
[en] is are , will have has must was should you
[es] es hay gracias mañana qué tiene tenemos trata lugar son
[fi] on ei ovat olisi oli toimitetaan eivät voi koskee voidaan
[fr] ’ ne est n' en a lieu aura vous avons
[it] , è che non si discussione svolgerà presidente ' sono
[nl] is zijn moeten heeft hebben moet kan zal wil wordt
[pt] , que é não senhor parlamento de aprova amanhã com
[sv] det jag vi detta vad därför de debatten ni den
```

5: noun (N)

```
[da] det dem her dette sig dag os noget betænkningen over
[de] abstimmung aussprache kommission bericht frage parlament
[el] θέμα έκθεση τροπολογία κ. Επιτροπή ψήφισμα πρόταση
[en] vote debate president commission much like council minutes
[es] debate presidente informe parlamento ello comisario sr.
[fi] unionin " yhteisön hetkellä enemmän uudelleen kerran heidän
[fr] vote débat parlement rapport commission question président
[it] votazione parlamento commissione relazione risoluzione
[nl] debat stemming commissie parlement verslag voorzitter
[pt] votação comissão questão relatório situação sessão proposta
[sv] det rum damer här detta oss sig frågan första dem
```

7: subject (SBJ)

```
[da] er har vil skal kan må var finder bør ville
[de] ich wir das , es sie – dies vielen (
[el] , θα δεν ψηφοφορία ότι Σώμα πολύ αυτό Πρόεδρε Το
[en] we i that it this there what ( thank they
[es] no se ¿ esto por lo me pero muchas tendrá
[fi] , että puhemies kiitoksia mietintö joka parlamentin mitä
[fr] nous je il c' cela j' ce l mais vous
[it] non ( la e questo ma si vorrei signor mi
[nl] dat ik wij het er we dit u daar wat
[pt] o , que encerrado presidente terá obrigado lugar isso não
[sv] är har måste kommer kan vill skall finns skulle var
```

9: preposition (PR)

```
[da] , af for i og til på med om fra
[de] und , in für auf von zu mit an auch
[el] και / ( σε , που από είναι των για
[en] to of in , for not and on with take
[es] , de que a en por con y para sobre
[fi] ja euroopan , on kuin / : tai ovat eikä
[fr] , de à que pour sur dans d' et par
[it] di e della del in a ( dell' dei da
[nl] van in , voor en op met aan over maar
[pt] de da do e para em dos / é com
[sv] i och för av till på om med som :
```

11: punctuation (.)

```
[da] . ? ) ! : f.eks. ... sessionen vedtoges bl.a.
[de] . ! ? ) : protokolls ... sitzungsperiode " ···
[el] . ) ! ; : κ. πρακτικών ... π.μ. κ.κ.
[en] . ? ) ! a.m. : p.m. ... ' ;
[es] . ? ) ! : ... ; » " anterior
[fi] . ? ) ! : ... " ···; ääntä
[fr] . ? ) ! la : ... » ; ···
[it] . ? ) ! : sessione ... " ; precedente
[nl] . ? ) ! : zitting ... gesloten " onderbroken
[pt] . ? ) ! : ... " anterior ; urgentes
[sv] . ! ) ? : protokoll ... sessionen t.ex. "
```

13: determiner (DT)

```
[da] ikke at en den også de et gerne det være
[de] die der eine den das ein diese im des dieser
[el] να το την η τη τις είναι τα στην μια
[en] the a be mr very an been not no in
[es] el ha este señor un se hemos debemos han debe
[fi] ole myös kuitenkin vielä hyvin siis erittäin nyt jo vain
[fr] le la les l' une cette un ce ces des
[it] la il l' un una le in i a gli
[nl] de het een deze dit geen die onze mijn mijnheer
[pt] a o uma um os de as esta este em
[sv] att inte en mycket också ett för vara om äga
```

Figure 2: Probable terminals of emission for each language and nonterminal.

0 → 16 11	(R → S .)	0.11
16 → 7 6	(S → SBJ VP)	0.06
6 → 2 12	(VP → V NP)	0.04
12 → 13 5	(NP → DT N)	0.19
15 → 17 19	(NP → NP N)	0.07
17 → 5 9	(NP → N PR)	0.07
15 → 13 5	(NP → DT N)	0.06

Figure 3: Examples of inferred common grammar rules in eleven languages, and their probabilities. Hand-provided annotations have the following meanings, R: root, S: sentence, NP: noun phrase, VP: verb phrase, and others appear in Figure 2.

We named nonterminals by using grammatical categories after the inference. We can see that words in the same grammatical category clustered across languages as well as within a language. Figure 3 shows examples of inferred common grammar rules with high probabilities. Grammar rules that seem to be common to European languages have been extracted.

5 Discussion

We have proposed a Bayesian hierarchical PCFG model for capturing commonalities at the syntax level for non-parallel multilingual corpora. Although our results have been encouraging, a number of directions remain in which we must extend our approach. First, we need to evaluate our model quantitatively using corpora with a greater diversity of languages. Measurement examples include the perplexity, and machine translation score. Second, we need to improve our model. For example, we can infer the number of nonterminals with a nonparametric Bayesian model (Liang et al., 2007), infer the model more robustly based on a Markov chain Monte Carlo inference (Johnson et al., 2007), and use probabilistic grammar models other than PCFGs. In our model, all the multilingual grammars are generated from a general model. We can extend it hierarchically using the coalescent (Kingman, 1982). That model may help to infer an evolutionary tree of languages in terms of grammatical structure without the etymological information that is generally used (Gray and Atkinson, 2003). Finally, the proposed approach may help to indicate the presence of a universal grammar (Chomsky, 1965), or to find it.

References

Phil Blunsom, Trevor Cohn, and Miles Osborne. 2009. Bayesian synchronous grammar induction. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 161–168.

Alexandre Bouchard-Côté, Percy Liang, Thomas Griffiths, and Dan Klein. 2008. A probabilistic approach to language change. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 169–176, Cambridge, MA. MIT Press.

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

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270, Morristown, NJ, USA. Association for Computational Linguistics.

Norm Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Shay B. Cohen and Noah A. Smith. 2009. Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 74–82, Morristown, NJ, USA. Association for Computational Linguistics.

Hal Daumé III. 2009. Bayesian multitask learning with latent hierarchies. In *Proceedings of the Twenty-Fifth Annual Conference on Uncertainty in Artificial Intelligence (UAI-09)*, pages 135–142, Corvallis, Oregon. AUAI Press.

Jason Eisner. 2003. Learning non-isomorphic tree mappings for machine translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 205–208, Morristown, NJ, USA. Association for Computational Linguistics.

Russell D. Gray and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426(6965):435–439, November.

Mark Johnson, Thomas Griffiths, and Sharon Goldwater. 2007. Bayesian inference for PCFGs via Markov chain Monte Carlo. In *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, Rochester, New York, April. Association for Computational Linguistics.

J. F. C. Kingman. 1982. The coalescent. *Stochastic Processes and their Applications*, 13:235–248.

Dan Klein and Christopher D. Manning. 2002. A generative constituent-context model for improved grammar induction. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 128–135, Morristown, NJ, USA. Association for Computational Linguistics.

Dan Klein and Christopher D. Manning. 2004. Corpus-based induction of syntactic structure: models of dependency and constituency. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 478, Morristown, NJ, USA. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86.

Kenichi Kurihara and Taisuke Sato. 2004. An application of the variational Bayesian approach to probabilistic context-free grammars. In *International Joint Conference on Natural Language Processing Workshop Beyond Shallow Analysis*.

K. Lari and S.J. Young. 1990. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer Speech and Language*, 4:35–56.

Percy Liang, Slav Petrov, Michael I. Jordan, and Dan Klein. 2007. The infinite PCFG using hierarchical dirichlet processes. In *EMNLP '07: Proceedings of the Empirical Methods on Natural Language Processing*, pages 688–697.

I. Dan Melamed. 2003. Multitext grammars and synchronous parsers. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 79–86, Morristown, NJ, USA. Association for Computational Linguistics.

Thomas Minka. 2000. Estimating a Dirichlet distribution. Technical report, M.I.T.

Michael P. Oakes. 2000. Computer estimation of vocabulary in a protolanguage from word lists in four daughter languages. *Journal of Quantitative Linguistics*, 7(3):233–243.

Steven Pinker. 1994. *The Language Instinct: How the Mind Creates Language*. HarperCollins, New York.

Benjamin Snyder, Tahira Naseem, and Regina Barzilay. 2009. Unsupervised multilingual grammar induction. 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*, pages 73–81, Suntec, Singapore, August. Association for Computational Linguistics.

Andreas Stolcke and Stephen M. Omohundro. 1994. Inducing probabilistic grammars by Bayesian model merging. In *ICGI '94: Proceedings of the Second International Colloquium on Grammatical Inference and Applications*, pages 106–118, London, UK. Springer-Verlag.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Comput. Linguist.*, 23(3):377–403.

Kai Yu, Volker Tresp, and Anton Schwaighofer. 2005. Learning gaussian processes from multiple tasks. In *ICML '05: Proceedings of the 22nd International Conference on Machine Learning*, pages 1012–1019, New York, NY, USA. ACM.