Labeled Grammar Induction with Minimal Supervision

Yonatan Bisk **Christos Christodoulopoulos** **Julia Hockenmaier**
Department of Computer Science
The University of Illinois at Urbana-Champaign
201 N. Goodwin Ave, Urbana, IL 61801
{bisk1,christod,juliahmr}@illinois.edu

Abstract

Nearly all work in unsupervised grammar induction aims to induce unlabeled dependency trees from gold part-of-speech-tagged text. These clean linguistic classes provide a very important, though unrealistic, inductive bias. Conversely, induced clusters are very noisy. We show here, for the first time, that very limited human supervision (three frequent words per cluster) may be required to induce *labeled* dependencies from automatically induced word clusters.

1 Introduction

Despite significant progress on inducing part-of-speech (POS) tags from raw text (Christodoulopoulos et al., 2010; Blunsom and Cohn, 2011) and a small number of notable exceptions (Seginer, 2007; Spitkovsky et al., 2011; Christodoulopoulos et al., 2012), most approaches to grammar induction or unsupervised parsing (Klein and Manning, 2004; Spitkovsky et al., 2013; Blunsom and Cohn, 2010) are based on the assumption that gold POS tags are available to the induction system. Although most approaches treat these POS tags as arbitrary, if relatively clean, clusters, it has also been shown that the linguistic knowledge implicit in these tags can be exploited in a more explicit fashion (Naseem et al., 2010). The presence of POS tags is also essential for approaches that aim to return richer structures than the standard unlabeled dependencies. Boonkwan and Steedman (2011) train a parser that uses a semi-automatically constructed Combinatory Categorial Grammar (CCG, Steedman (2000)) lexicon for POS tags, while Bisk and Hockenmaier (2012; 2013) show that CCG lexicons can be induced automatically if POS tags are used to identify nouns and verbs. However, assuming clean POS tags is highly unrealistic for most scenarios in which one would wish to use an otherwise unsupervised parser.

In this paper we demonstrate that the simple "universal" knowledge of Bisk and Hockenmaier (2013) can be easily applied to induced clusters given a small number of words labeled as *noun*, *verb* or *other*, and that this small amount of knowledge is sufficient to produce labeled syntactic structures from raw text, something that has not yet been proposed in the literature. Specifically, we will provide a labeled evaluation of induced CCG parsers against the English (Hockenmaier and Steedman, 2007) and Chinese (Tse, 2013) CCGbanks. To provide a direct comparison to the dependency induction literature, we will also provide an unlabeled evaluation on the 10 dependency corpora that were used for the task of grammar induction from raw text in the PASCAL Challenge on Grammar Induction (Gelling et al., 2012).

The system of Christodoulopoulos et al. (2012) was the only participant competing in the PASCAL Challenge that operated over raw text (instead of gold POS tags). However, their approach did not outperform the six baseline systems provided. These baselines were two versions of the DMV model (Klein and Manning, 2004; Gillenwater et al., 2011) run on varying numbers of induced Brown clusters (described in section 2.1). We will therefore compare against these baselines in our evaluation.

Outside of the shared task, Spitkovsky et al. (2011) demonstrated impressive performance using Brown clusters but did not provide evaluation

for languages other than English.

The system we propose here will use a coarse-grained labeling comprised of three classes, which makes it substantially simpler than traditional tagsets, and uses far fewer labeled tokens than is customary for weakly-supervised approaches (Haghighi and Klein, 2006; Garrette et al., 2015).

2 Our Models

Our goal in this work will be to produce labeled dependencies from raw text. Our approach is based on the HDP-CCG parser of Bisk and Hockenmaier (2015) with their extensions to capture lexicalization and punctuation, which, to our knowledge, is the only unsupervised approach to produce labeled dependencies. It first induces a CCG from POS-tagged text, and then estimates a model based on Hierarchical Dirichlet Processes (Teh et al., 2006) over the induced parse forests. The HDP model uses a hyperparameter which controls the amount of smoothing to the base measure of the HDP. Setting this value will prove important when moving between datasets of drastically different sizes.

The induction algorithm assumes that a) verbs may be predicates (with category S), b) verbs can take nouns (with category N) or sentences as arguments (leading to categories of the form S|N, (S|N)|N, (S|N)|S etc.), c) any word can act as a modifier, i.e. have a category of the form X|X if it is adjacent to a word with category X or X|Y, and d) modifiers X|X can take nouns or sentences as arguments ((X|X)|N). Our contribution in this paper will be to show that we can replace the gold POS tags used by Bisk and Hockenmaier (2013) with automatically induced word clusters, and then use very minimal supervision to identify noun and verb clusters.

2.1 Inducing Word Clusters

We will evaluate three clustering approaches:

Brown Clusters Brown clusters (Brown et al., 1992) assign each word to a single cluster using an aglomerative clustering that maximizes the probability of the corpus under a bigram class conditional model. We use Liang's implementation[1].

BMMM The Bayesian Multinomial Mixture Model[2] (BMMM, Christodoulopoulos et al. 2011) is also a hard clustering system, but has the ability to incorporate multiple types of features either at a token level (e.g. ± 1 context word) or at a type level (e.g. morphology features derived from the Morfessor system (Creutz and Lagus, 2006)). The combination of these features allows BMMM to better capture morphosyntactic information.

Bigram HMM We also evaluate unsupervised bigram HMMs, since the soft clustering they provide may be advantageous over the hard Brown and BMMM clusters. But it is known that unsupervised HMMs may not find good POS tags (Johnson, 2007), and in future work, more sophisticated models (e.g. Blunsom and Cohn (2011)), might outperform the systems we use here.

In all cases, we assume that we can identify punctuation marks, which are moved to their own cluster and ignored for the purposes of tagging and parsing evaluation.

2.2 Identifying Noun and Verb Clusters

To induce CCGs from induced clusters, we need to label them as $\{noun, verb, other\}$. This needs to be done judiciously; providing every cluster the *verb* label, for example, leads to the model identifying prepositions as the main sentential predicates.

We demonstrate here that labeling three frequent words per cluster is sufficient to outperform state-of-the-art performance on grammar induction from raw text in many languages. We emulate having a native speaker annotate words for us by using the universal tagset (Petrov et al., 2012) as our source of labels for the most frequent three words per cluster (we map the tags NOUN, NUM, PRON to *noun*, VERB to *verb*, and all others to *other*). The final labeling is a majority vote, where each word type contributes a vote for each label it can take (see Table 4 for some examples). This approach could easily be scaled to allow more words per cluster to vote. But we will see that three per cluster is sufficient to label most tokens correctly.

3 Experimental Setup

We will focus first on producing CCG labeled predicate-argument dependencies for English and Chinese and will then apply our best settings to produce a comparison with the tree structures of the languages of the PASCAL Shared Task. All languages will be trained on sentences of up to length 20 (not counting punctuation). All cluster induction algorithms are treated as black boxes

[1] https://github.com/percyliang/brown-cluster

[2] https://github.com/christos-c/bmmm

and run over the complete datasets in advance. This alleviates having to handle tagging of unknown words.

To provide an intuition for the performance of the induced word clusters, we provide two standard metrics for unsupervised tagging:

Many-to-one (M-1) A commonly used measure, M-1 relies on mapping each cluster to the most common POS tag of its words. However, M-1 can be easily inflated by inducing more clusters.

V-Measure Proposed by Rosenberg and Hirschberg (2007), V-Measure (VM) measures the information-theoretic distance between two clusterings and has been shown to be robust to the number of induced clusters (Christodoulopoulos et al., 2010). Both of these metrics are known to be highly dependent on the gold annotation standards they are compared against, and may not correlate with downstream performance at parsing.

Of more immediate relevance to our task is the ability to accurately identify nouns and verbs:

Noun, Verb, and Other Recall We measure the (token-based) recall of our three-way labeling scheme of clusters as *noun/verb/other* against the universal POS tags of each token.

4 Experiment 1: CCG-based Evaluation

Experimental Setup For our primary experiments, we train and test our systems on the English and Chinese CCGbanks, and report directed labeled F1 (LF1) and undirected unlabeled F1 (UF1) over CCG dependencies (Clark et al., 2002). For the labeled evaluation, we follow the simplification of CCGbank categories proposed by Bisk and Hockenmaier (2015): for English to remove morphosyntactic features, map NP to N and change VP modifiers (S\NP)|(S\NP) to sentential modifiers (S|S); for Chinese we map both M and QP to N. In the CCG literature, UF1 is commonly used because undirected dependencies do not penalize argument vs. adjunct distinctions, e.g. for prepositional phrases. For this reason we will include UF1 in the final test set evaluation (Table 2).

We use the published train/dev/test splits, using the dev set for choosing a cluster induction algorithm, and then present final performance on the test data. We induce 36 tags for English and 37 for Chinese to match the number of tags present in the treebanks (excluding symbol and punctuation tags).

		Tagging		Labeling			Parsing	
		M-1	VM	N	/ V	/ O	LF1	Gold
English	Brown	62.4	56.3	**85.6**	59.4	81.2	23.3	
	BMMM	**66.8**	**58.7**	81.0	**81.2**	**82.7**	**26.6**	38.8
	HMM	51.1	41.7	76.3	63.3	**82.6**	25.8	
Chinese	Brown	**66.0**	**50.1**	88.9	28.6	**91.3**	10.2	
	BMMM	64.8	50.0	**94.4**	**48.7**	87.0	**10.5**	16.6
	HMM	46.3	30.8	68.0	44.6	76.7	3.13	

Table 1: Tagging evaluation (M-1, VM, N/V/O Recall) and directed labeled CCG-Dependency performance (LF1) as compared to the use of gold POS tags (Gold) for three clustering algorithms.

Results Table 1 presents the parsing and tagging development results on the two CCG corpora. In terms of tagging performance, we can see that the two hard clustering systems significantly outperform the HMM, but the relative performance of Brown and BMMM is mixed.

More importantly, we see that, at least for English, despite clear differences in tagging performance, the parsing results (LF1) are much more similar. In Chinese, we see that the performance of the two hard clustering systems is almost identical, again, not representative of the differences in the tagging scores. The N/V/O recall scores in both languages are equally poor predictors of parsing performance. However, these scores show that having only three labeled tokens per class is sufficient to capture most of the necessary distinctions for the HDP-CCG. All of this confirms the observations of Headden et al. (2008) that POS tagging metrics are not correlated with parsing performance. However, since BMMM seems to have a slight overall advantage, we will be using it as our clustering system for the remaining experiments.

Since the goal of this work was to produce labeled syntactic structures, we also wanted to evaluate our performance against that of the HDP-CCG system that uses gold-standard POS tags. As we can see in the last two columns of our development results in Table 1 and in the final test results of Table 2, our system is within 2/3 of the labeled performance of the gold-POS-based HDP-CCG[3].

Figure 1 shows an example labeled syntactic structure induced by the model. We can see the system successfully learns to attach the final

[3]To put this result into its full perspective, the LF1 performance of a supervised CCG system (Hockenmaier and Steedman, 2002), HWDep model, trained on the same length-20 dataset and tested on the simplified CCGbank test set is 80.3.

	This	Gold
English	26.0 / 51.1	37.1 / 64.9
Chinese	10.3 / 33.5	15.6 / 39.8

Table 2: CCG parsing performance (LF1/UF1) on the test set with and without gold tags.

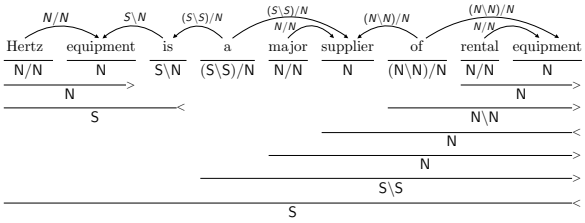

Figure 1: A sample derivation from the WSJ Section 22 demonstrating the system is learning most of the correct categories of CCGbank but has incorrectly analyzed the determiner as a preposition.

prepositional phrase, but mistakes the verb for intransitive and treats the determiner *a* as a preposition. The labeled and undirected recall for this parse are 5/8 and 7/8 respectively.

5 Experiment 2: PASCAL Shared Task

Experimental Setup During the PASCAL shared task, participants were encouraged to train over the complete union of the data splits. We do the same here, use the dev set for choosing a HDP-CCG hyperparameter, and then present final results for comparison on the test section. We vary the hyperparamter for this evaluation because the datasets fluctuate dramatically in size from 9K to 700K tokens on sentences up to length 20. Rather than match all of the tagsets, we simply induce 49 (excluding punctuation) classes for every language. The actual tagsets vary from 20 to 304 tags (median 39, mean 78).

Results We now present results for the 10 corpora of the PASCAL shared task (evaluated on all sentence lengths). Table 3 presents the test performance for each language with the best hyperparameter chosen from the set $\{100, 1000, 2500\}$. Also included are the best published results from the joint tag/dependency induction shared task (ST) as well as the results from Bisk and Hockenmaier (2013), the only existing numbers for multilingual CCG induction (BH) with gold part-of-speech tags. Note that the systems in ST do not have access to any gold-standard POS tags, whereas our system has access to the gold tags for the three most frequent words of each cluster.

	VM	N / V / O	This	ST	@15	BH
Czech_{2500}	42	86 / 67 / 67	9.49	**33.2**	12.2	50.7
English_{2500}	59	87 / 76 / 85	**43.8**	24.4	51.6	62.9
CHILDES_{2500}	68	84 / 97 / 89	**47.2**	42.2	47.5	73.3
Portuguese_{2500}	55	88 / 81 / 69	**55.5**	31.7	55.8	70.5
Dutch_{1000}	50	81 / 81 / 82	**39.9**	33.7	43.8	54.4
Basque_{1000}	52	2 / 78 / 95	**31.1**	28.7	35.2	45.0
Swedish_{1000}	50	89 / 74 / 85	**45.8**	28.2	52.9	66.9
Slovene_{1000}	50	83 / 75 / 79	18.5	**19.2**	23.6	46.4
Danish_{100}	59	95 / 79 / 82	16.1	**31.9**	17.8	58.5
Arabic_{100}	51	85 / 76 / 90	34.5	**44.4**	43.7	65.1
Average	54	78 / 78 / 82	**34.2**	31.8	38.4	59.4

Table 3: Tagging VM and N/V/O Recall alongside Directed Accuracy for our approach and the best shared task baseline. Additionally, we provide results for length 15 to compare to previously published results ([ST]: Best of the PASCAL joint tag/dependency induction shared task systems; [BH]: Bisk and Hockenmaier (2013).

The languages are sorted by the number of non-punctuation tokens in sentences of up to length 20. Despite our average performance (34.2) being slightly higher than the shared task (31.8), the st. deviation is substantial ($\sigma = 15.2$ vs $\sigma_{ST} = 7.5$). It seems apparent from the results that while data sparsity may play a role in affecting performance, the more linguistically interesting thread appears to be morphology. Czech is perhaps a prime example, as it has twice the data of the next largest language (700K tokens vs 336K in English), but our approach still performs poorly.

Finally, while we saw that the hard clustering systems outperformed the HMM for our experiments, this is perhaps best explained by analyzing the average number of gold fine-grained tags per lexical type in each of the corpora. We found, counterintuitively, that the "difficult" languages had lower average number of tags per type (1.01 for Czech, 1.03 for Arabic) than English (1.17) which was the most ambiguous. This is likely due to morphology distinguishing otherwise ambiguous lemmas.

6 Cluster Analysis

In Table 4, we present the three most frequent words from several clusters produced by the BMMM for English and Chinese. We also provide a *noun*/*verb*/*other* label for each of the words in the list. One can clearly see that there are many ambiguous cases where having three labels voting

English	Labels	Chinese	Chinese gloss	Labels
shares, sales, business	N, N, N	同时, 政治, 生产	simultaneously, politics, production	O, N, N
the, its, their	O, N, N	进行, 举行, 开始	advance, hold, begin	V, V, V
other, interest, chief	O, N, O	在, 有, 对	in, have, for	O, V, O
of, in, on	O, O, O	中国, 台湾, 美国	China, Taiwan, USA	N, N, N
up, expected, made	O, V, V	也, 将, 就	also, will, then	O, O, O
be, make, sell	V, V, V	大, 多,高	big, many, high	O, N, O *
offer, issue, work	N, N, N *	是, 希望, 代表	is, desire, representative	V, V, N

Table 4: The top three words in BMMM clusters with their *noun*/*verb*/*other* labels. In two cases (marked with *) all three of the most frequent words also occurred as a verb at least one third of the time.

on the class label proves a beneficial signal. We have also marked two classes with * to draw the reader's attention to a fully *noun* cluster in English and an *other* cluster in Chinese which are highly ambiguous. Specifically, in both of these cases the frequent words also occur frequently as verbs, providing additional motivation for a better soft-clustering algorithm in future work.

How to most effectively use seed knowledge and annotation is still an open question. Approaches range from labeling frequent words like the work of Garrette and Baldridge (2013) to the recently introduced active learning approach of Stratos and Collins (2015). In this work, we were able to demonstrate high noun and verb recall with the use of a very small set of labeled words because they correspond to an existing clustering. In contrast, we found that labeling even the 1000 most frequent words led to very few clusters being correctly identified; e.g. in English, using the 1000 most frequent words results in identifying 2 *verb* and 5 *noun* clusters, compared to our method's 9 *verb* and 16 *noun* clusters. This is because the most frequent words tend to be clustered in a few very large clusters resulting in low coverage.

Stratos and Collins (2015) demonstrated, similarly, that using a POS tagger's confidence score to find ambiguous classes can lead to a highly effective adaptive learning procedure, which strategically labels very few words for a very highly accurate system. Our results align with this research, leading us to believe that this paradigm of guided minimal supervision is a fruitful direction for future work.

7 Conclusions

In this paper, we have produced the first labeled syntactic structures from raw text. There remains a noticeable performance gap due to the use of induced clusters in lieu of gold tags. Based on our final PASCAL results, there are several languages where our performance greatly exceeds the currently published results, but equally many where we fall short. It also appears to be the case that this problem correlates with morphology (e.g. Arabic, Danish, Slovene, Basque, Czech) and some of the lowest performing intrinsic evaluations of the clustering and N/V/O labeling (Czech and Basque).

In principle, the BMMM is taking morphological information into account, as it is provided with the automatically produced suffixes of Morfessor. Unfortunately, its treatment of them simply as features from a "black box" appears to be too naive for our purposes. Properly modeling the relationship between prefixes, stems and suffixes both within the tag induction and parsing framework is likely necessary for a high performing system.

Moving forward, additional raw text for training, as well as enriching the clustering with induced syntactic information (Christodoulopoulos et al., 2012) may close this gap.

8 Acknowledgments

We want to thank Dan Roth and Cynthia Fisher for their insight on the task. Additionally, we would like to thank the anonymous reviewers for their useful questions and comments. This material is based upon work supported by the National Science Foundation under Grants No. 1053856, 1205627, 1405883, by the National Institutes of Health under Grant HD054448, and by DARPA under agreement number FA8750-13-2-0008. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the National Institutes of Health, DARPA or the U.S. Government.

References

Yonatan Bisk and Julia Hockenmaier. 2012. Simple Robust Grammar Induction with Combinatory Categorial Grammars. In *Proceedings of the Twenty-Sixth Conference on Artificial Intelligence (AAAI-12)*, pages 1643–1649, Toronto, Canada, July.

Yonatan Bisk and Julia Hockenmaier. 2013. An HDP Model for Inducing Combinatory Categorial Grammars. *Transactions of the Association for Computational Linguistics*, pages 75–88.

Yonatan Bisk and Julia Hockenmaier. 2015. Probing the linguistic strengths and limitations of unsupervised grammar induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*.

Phil Blunsom and Trevor Cohn. 2010. Unsupervised Induction of Tree Substitution Grammars for Dependency Parsing. *Proceedings of the 2010 Conference on Empirical Methods of Natural Language Processing*, pages 1204–1213, October.

Phil Blunsom and Trevor Cohn. 2011. A hierarchical pitman-yor process hmm for unsupervised part of speech induction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 865–874, Portland, Oregon, USA, June.

Prachya Boonkwan and Mark Steedman. 2011. Grammar Induction from Text Using Small Syntactic Prototypes. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 438–446, Chiang Mai, Thailand, November.

Peter F Brown, Peter V deSouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai. 1992. Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2010. Two decades of unsupervised PoS induction: How far have we come? In *Proceedings of EMNLP*, pages 575–584.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2011. A Bayesian Mixture Model for Part-of-Speech Induction Using Multiple Features. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK., July.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2012. Turning the pipeline into a loop: iterated unsupervised dependency parsing and PoS induction. In *WILS '12: Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, June.

Stephen Clark, Julia Hockenmaier, and Mark Steedman. 2002. Building deep dependency structures using a wide-coverage ccg parser. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 327–334, Philadelphia, Pennsylvania, USA, July.

Mathias Creutz and Krista Lagus. 2006. Morfessor in the Morpho challenge. In *Proceedings of the PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes*, pages 12–17.

Dan Garrette and Jason Baldridge. 2013. Learning a Part-of-Speech Tagger from Two Hours of Annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 138–147, Atlanta, Georgia, June.

Dan Garrette, Chris Dyer, Jason Baldridge, and Noah A Smith. 2015. Weakly-Supervised Grammar-Informed Bayesian CCG Parser Learning. In *Proceedings of the Association for the Advancement of Artificial Intelligence*.

Douwe Gelling, Trevor Cohn, Phil Blunsom, and João V Graca. 2012. The PASCAL Challenge on Grammar Induction. In *NAACL HLT Workshop on Induction of Linguistic Structure*, pages 64–80, Montréal, Canada, June.

Jennifer Gillenwater, Kuzman Ganchev, João V Graca, Fernando Pereira, and Ben Taskar. 2011. Posterior Sparsity in Unsupervised Dependency Parsing. *The Journal of Machine Learning Research*, 12:455–490, February.

Aria Haghighi and Dan Klein. 2006. Prototype-Driven Grammar Induction. In *Association for Computational Linguistics*, pages 881–888, Morristown, NJ, USA.

William P. Headden, III, David McClosky, and Eugene Charniak. 2008. Evaluating unsupervised part-of-speech tagging for grammar induction. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 329–336, Stroudsburg, PA, USA.

Julia Hockenmaier and Mark Steedman. 2002. Generative models for statistical parsing with combinatory categorial grammar. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 335–342, Philadelphia, Pennsylvania, USA, July.

Julia Hockenmaier and Mark Steedman. 2007. CCGbank: A Corpus of CCG Derivations and Dependency Structures Extracted from the Penn Treebank. *Computational Linguistics*, 33:355–396, September.

Mark Johnson. 2007. Why doesn't EM find good HMM POS-taggers. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, January.

Dan Klein and Christopher D Manning. 2004. Corpus-Based Induction of Syntactic Structure: Models of Dependency and Constituency. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 478–485, Barcelona, Spain, July.

Tahira Naseem, Harr Chen, Regina Barzilay, and Mark Johnson. 2010. Using universal linguistic knowledge to guide grammar induction. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1234–1244, Cambridge, MA, October.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A Universal Part-of-Speech Tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2089–2096, Istanbul, Turkey, May.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic, June.

Yoav Seginer. 2007. Fast Unsupervised Incremental Parsing. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 384–391, Prague, Czech Republic, June.

Valentin I Spitkovsky, Hiyan Alshawi, Angel X Chang, and Daniel Jurafsky. 2011. Unsupervised Dependency Parsing without Gold Part-of-Speech Tags. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1281–1290, Edinburgh, Scotland, UK., July.

Valentin I Spitkovsky, Hiyan Alshawi, and Daniel Jurafsky. 2013. Breaking Out of Local Optima with Count Transforms and Model Recombination: A Study in Grammar Induction. In *Empirical Methods in Natural Language Processing*.

Mark Steedman. 2000. *The Syntactic Process*. The MIT Press, September.

Karl Stratos and Michael Collins. 2015. Simple semi-supervised pos tagging. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 79–87, Denver, Colorado, June.

Yee-Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Daniel Tse. 2013. *Chinese CCGBank: Deep Derivations and Dependencies for Chinese CCG Parsing*. Ph.D. thesis, The University of Sydney.