

The position at CMU that I was hired to fill sought “statistical parsing” expertise. Since then, my interests have expanded quite a bit beyond statistical parsing, in no small part due to the stimulating environment at CMU. Here I will aim to summarize what I have accomplished and my trajectory in research, service, and educational activities.

1 Research

I am a **natural language processing** (NLP) researcher; I seek algorithms for automated understanding of text. My research program has three emphases:

1. Computational linguistics: How can natural language text and its structure be modeled computationally, capitalizing on properties that are universal across languages, while allowing for variation?
2. Machine learning for language: To what extent can such models be acquired *empirically* from either (i) data annotated by linguistic experts or (ii) raw text data?
3. Applications: How can the design and acquisition of such computational models be driven to support real applications that will help people?

Recurring themes in my broad research program are general problems arising across the field (e.g., structure prediction) and the integration of linguistic representations with data, and of computational models with applications. Below, I will discuss my work on each of the three questions above, each in turn.

1.1 Computational Linguistics

In computational linguistics, I have been a leader in drawing connections between *formalisms* (particularly syntactic and weighted formalisms) and practical problems in linguistic analysis.

Before CMU. As a graduate student, my research focused largely on models of syntax and algorithms for parsing [1, 2, 3, 4, 5], but also morphological disambiguation [6]. I was particularly interested in developing parsing methods adaptable for text in a variety of languages, since I viewed machine translation as the primary beneficiary of advances in computational linguistics [7, 8, 9]. I also developed an interest in generic algorithms for text processing, and was a major contributor to and user of Jason Eisner’s Dyna language for implementing dynamic programming algorithms [10, 11]. Dynamic programming is the most common starting point for implementing linguistic analyzers.

Current research. Over time, I have gradually moved from describing this research as “statistical parsing” to “linguistic structure prediction.” (Indeed, *Linguistic Structure Prediction* is the title of my forthcoming book [12].) This terminology shift highlights some important trends: (i) many linguistic analysis problems are better solved in formalisms other than parsing/dynamic programming, which relies on strong independence assumptions; (ii) the usefulness of “structure(d) prediction” techniques developed in machine learning, for linguistic problems; and (iii) the development of these ideas for natural language representations beyond syntax.

In both research and teaching, I advocate high-level representations of complex models and algorithms, for improved efficiency, modularity, understandability, and ease of implementation. This is especially important as we seek to model richer linguistic phenomena. Building on my earlier work in the weighted logic programming formalism for dynamic programming [11], my students and I have been exploring how automatic program transformations can be used to automate design and implementation of complicated multi-structure

NLP algorithms [13, 14]. An example is the derivation of bilingual parsing algorithms, as in [2] and now widely used in statistical machine translation [15], from simpler monolingual ones.

I have worked in other formalisms as well:

- A study of the expressive power of different kinds of weighted context-free and finite-state grammars [16];
- A study of sample complexity for probabilistic grammars (how much data is required for accurate learning of a grammar?) [17];
- Approximations for inference about linguistic structure when optimal substructure assumptions make dynamic programming unavailable, even though it suggests a partial solution [18, 19];
- Efficient polynomial-sized encodings of NP-hard linguistic structure prediction problems as integer linear programs [20];
- Novel inference algorithms for nonprojective dependency syntax, based on generalizations of Kirchhoff's matrix-tree theorem [21] and linear-time inference for dependency syntax models with hard constraints [22];
- The use of linguistic structure prediction for more global reasoning in problems outside syntax, like frame-semantic analysis [23, 24] and sentence-to-sentence transformations [25, 26]; and
- Connections between information theoretic measures and kernels for structured data like text [27, 28], and the use of kernel learning as a principled way to select features for linguistic structure prediction problems [29].

This line of research has been supported by a two-year DARPA grant (NBCH-1080004) and a one-year NSF grant (NSF-IIS-0713265). Two of my Ph.D. students, Dipanjan Das (expected to graduate in 2011) are focusing on computational linguistics research.

Future directions. Ongoing CL/NLP research will focus on formalisms that support linguistic analysis with richer and more non-local features. I am especially interested in models that make use of implicit feature spaces (e.g., through kernelization) and symbolic latent variables. Both of these pose algorithmic challenges for prediction.

Recently I was awarded a grant, with co-PI Kemal Oflazer, by the Qatar National Research Foundation (NPRP-08-485-1-083) to study cross-lingual models for improving NLP for Arabic. Integrating multiple levels of linguistic structure (e.g., morphology and syntax) is one such challenge that we consider in this research project. Modeling the relationship between structure in English (which can be predicted noisily) and structure in another language like Arabic (for which performance is much worse) is a longstanding challenge in NLP, and we plan to make advances through the use of probabilistic structure prediction models.

The development of models that are applicable across languages, and the use of accurate processing for resource-rich languages (like English) to improve processing in other languages (like Arabic) remains a long-term interest.

1.2 Machine Learning for Language

In machine learning, I have made leading contributions to unsupervised learning of structure, in particular developing principled, efficient, and effective techniques for combining domain knowledge (e.g., linguistics) with statistical learning.

Before CMU. My dissertation was about unsupervised learning of natural language syntax [30]: can a learner discover the syntactic structure of a language from unparsed sentences in that language? Noting that most unsupervised language learning begins with the idea of maximizing likelihood, I developed techniques to overcome (i) inherent problems with likelihood as an objective, proposing a generalized class of alternatives I called “contrastive estimation” [3] and (ii) the difficulty of the optimization problem itself, in both the likelihood and contrastive cases, proposing variations on deterministic annealing to avoid local optima [1, 5]. A recurring theme in my graduate research was drawing tighter connections between supervised and unsupervised learning.

Current research. Like many LTI faculty, my secondary appointment at CMU is in the Machine Learning Department. My machine learning research is focused on advancing machine learning to be able to handle the hard problems presented by natural language, including: intractability of inference, complex evaluation criteria, and the expense of generating high-quality annotated data.

“Inference” can refer to making a prediction (given a linguistic input x , return the model’s favorite output y), but also to the calculation of marginals and other statistics. Inference algorithms are applied within the final language processing application, and are also usually found at the innermost loop of learning algorithms, so their runtime efficiency is very important. Traditional models of linguistic structure accomplished exact inference using dynamic programming, but (as discussed above), today’s models are more expensive because they make weaker independence assumptions. My collaborators and I have studied how *approximations* in inference interact with learning [31], and also approximate learning methods like stacking [32] and novel approaches to learning that avoid inference altogether [33]. My students and I have also developed techniques for using “clouds” and other multiprocessor platforms to perform learning in a distributed setting [34].

Another major challenge of natural language processing is *evaluation*. NLP researchers have developed a wide range of techniques for measuring the quality of an NLP component. Such criteria should be taken into consideration when applying machine learning to linguistic data, and discriminative techniques seek to do exactly that. Recently my students and I have developed new learning objectives that enable the direct inclusion of cost functions (evaluation scores) in structured prediction models without sacrificing convexity or the useful foundation of probabilistic modeling [35]. We have also developed new fast-converging online learning algorithms for such models that generalize much previous work and do not require tuning a learning rate [36].

As noted, my graduate research focused on unsupervised learning of syntactic structure. Since my dissertation, unsupervised dependency parsing has become an extremely popular topic. I have continued to make leading contributions in this area, including the application of Bayesian modeling to grammar induction and related unsupervised learning problems like segmentation, in parametric [37, 38, 39, 40] and nonparametric [41, 42] settings. Recently my students and I have also made theoretical advances, proving new NP-hardness results for certain unsupervised grammar learning problems [43].

This research is supported by a three-year NSF grant (IIS-0915187), a one-year NSF SGER (IIS-0836431), and a CMU-Portugal joint Ph.D. fellowship to André Martins. Two of my Ph.D. students, André Martins and Shay Cohen, are expected to graduate (both in 2011) as machine learning researchers with an emphasis on learning structure. Chris Dyer from the University of Maryland will join my group in August 2010 as a postdoctoral researcher focusing on unsupervised learning for NLP.

Future directions. An area of continued interest is learning in the face of approximate inference. I am particularly interested in variational techniques, but these are challenging from an engineering perspective. Future work will develop generic methods for automatically deriving variational updates, to support a wide range of structured modeling problems. Another approach that has been under-explored for structured problems is *importance sampling* (e.g., particle filtering for HMMs). This approach may have runtime advantages

compared to more widely used MCMC methods, and may also be easier to understand and implement.

Another major direction of interest is the use of linguistic bias in unsupervised learning. All unsupervised learners capitalize on bias, but only rarely is this bias formulated in a way that is intuitive to people who understand the data. The vast majority of machine learning research in this space is motivated more by the computational ease of inference (even approximate inference) rather than by the peculiarities of the linguistic domain. I intend to continue developing techniques for encoding bias, e.g., through Bayesian priors and regularization methods, that will make statistical language learning more amenable to combination with linguistics.

Finally, I have recently begun discussions and proposal-writing with researchers in computer vision, both at CMU and outside CMU, to begin exploring how my research in unsupervised structure discovery can be explored in problems of modeling human activity as seen in visual data.

1.3 Applications

I hold the view that AI research must be motivated by concrete applications or risk falling into the cracks between CS theory and engineering. This does not mean that AI researchers need to *produce* commercial-strength implementations; I think that industry is usually far better suited to that role. Academics may be able to identify *new* applications and to discover new algorithms to improve existing applications; these are where my focus has been. Currently I am the initiator of the application of NLP to *forecasting*, a new challenge with potential for high impact.

Before CMU. My initial research experiences involved working on symbolic and statistical machine translation projects as an undergraduate (at the University of Maryland in 1998 and at the Johns Hopkins University CLSP summer workshop in 1999). As a graduate student my interests turned mainly to more formal problems (aside from a summer working on personalized information retrieval at Microsoft Research in 2004).

Current research. I have a continued interest in machine translation, and have collaborated with many MT researchers at CMU on making better use of parallel data and linguistic analysis in building MT systems [44, 45, 15]. Another contribution is the development of a new translation system based on quasi-synchronous grammars, which use syntax but do not constrain source and target structures to be isomorphic [46]. Other widely studied NLP applications include document summarization and question answering; my collaborators and I have developed and evaluated models for parts of these problems: sentence compression [47], joint compression and sentence selection [48], and answer selection for question answering [49].

An application area receiving much attention at CMU is the field of education technologies. One of my Ph.D. students, Michael Heilman, started out in this area and began to collaborate with me to advance linguistic models for use in educational problems. His dissertation research focuses on automatic generation of questions, for testing comprehension of reading materials [50, 51, 52, 53]. This is a natural language generation problem, but the input is text. Heilman and I have developed a linguistically-informed framework that produces fact-oriented questions from an arbitrary piece of text.

A more open direction is the development of applications based on social media like blogs and microblogs. My collaborators and I have explored political blogs in particular, building predictive models of the discourse found in comments on political blogs [54, 55]. We have also begun to explore ways of measuring *bias* in political blogs [56], and shown temporal correlation between sentiment-word statistics in large volumes of microblog posts and traditional public opinion polls [57].¹ While many research groups have

¹This work recently attracted the attention of the international news media and blogosphere, including CNN, the BBC, Huffington Post, SlashDot, and others.

built models of the blogosphere’s social structure, we are among the few who have considered the *language* of this relatively new text domain.

An exciting new direction where I have taken a leading role is text-driven *forecasting*, where text inputs are used to make concrete predictions about the future. I have developed regression models that take shallow text statistics as input, and shown that they can make accurate predictions about financial risk [58] and opening-weekend revenues of films [59]. To date we have used simple and shallow techniques and achieved very promising results on a few problems, and I believe further work will show that many kinds of concrete real world events and measurements are predictable from relevant text collections, especially social media.

This research has been primarily supported by industry gifts, an NSF fellowship to Michael Heilman, and a two-year NSF grant to Stephan Vogel and myself (IIS-0844507). Three of my Ph.D. students are expected to graduate with dissertations on text applications: Michael Heilman (question generation, in 2011), Kevin Gimpel (machine translation, in 2011), and Tae Yano (social media).

Future directions. I see forecasting as a major thrust of my future work. CMU is in negotiation with IARPA for a grant that will support the development of forecasting of text content in future scientific articles, by modeling trends in research literature. I also see forecasting as an exciting application that can leverage social media, particularly in the political domain. This spring and summer I am spending considerable time at interdisciplinary meetings with political scientists to identify research directions that will aid text-driven social science research. I have a continued, close collaboration with Bryan Routledge of the Tepper School of Business at CMU.

I have begun collaborating with Tom Mitchell and William Cohen on the new Google-sponsored “Worldly Knowledge” project. I would like to see a tighter connection develop between structured linguistic analysis and structured information extraction. How can NLP contribute to enriching knowledge bases, and how can knowledge bases, text, and non-text data (as in the forecasting scenario) be modeled jointly? This is a challenging application area that builds on considerable disparate past work in NLP and machine learning.

1.4 Summary

My research interests are broad, but I believe they are mutually reinforcing. The problems that arise in NLP do not usually admit easy machine learning solutions. Machine learning algorithms will not be adopted if they are not validated on real tasks like those that arise in NLP. To be a modern NLP researcher—simply to read the latest papers—requires fluency in both CL and ML. While these two communities have different styles and aesthetics, I believe it is a necessary investment to actively participate in both.

Further, motivating theoretical and formal work in computational linguistics and machine learning by real applications gives the best chances of high impact within and beyond the research community. Hence I try to stay “grounded,” and contribute to the field by identifying new application challenges (e.g., forecasting) that can drive our research forward.

2 Education

This statement describes my educational philosophy through a discussion of advising and teaching.

2.1 Advising and Leading a Research Group

As a graduate advisor, it is my duty and privilege to help students turn themselves into scholars. This is the most important part of my job and, if I am successful, will have the greatest impact in the long term. Hence it receives more of my time than any other activity.

At this writing I advise seven Ph.D. students (two co-advised), each of whom I meet one-on-one for an hour a week. Since coming to CMU I have mentored more than fifteen additional students (graduates, undergraduates, and recent graduates I've employed as programmers) through regular, one-on-one weekly hour meetings over a period of at least a semester.

My goal is to train my students not only to independently conduct novel, well-motivated, and high-impact formal and experimental research, but also to communicate effectively to the scientific and broader community, understand where their work fits in the landscape of the field, and develop the skills to critically appraise research, including their own. As soon as I began advising students (my first month as a professor) I instituted a weekly research group meeting to build a research culture of collectivism where everyone in the group benefits from the insights of the rest. My group meetings (now 90 minutes a week, 120 in the summer) are lively and typically delve into deep technical details of our work and papers we read together. This is especially helpful in painting a “big picture,” given the broad range of interests in the group. For undergraduates who work in my group, attending group meetings offers exposure to a fast-paced and broad-based research process, giving them a taste of graduate school.

The benefits of a cohesive research group include informal within-group mentoring and a steady flow of multiple-student-authored papers. I encourage collaboration outside the group, as well, as the group's publication record shows.

Looking ahead. My research group is now about ten researchers (students, programmer, postdoc), plus undergraduates. I believe this is about as large as the group can grow without additional faculty hiring in the area of statistical natural language processing. Until then, I hope to maintain a steady state and continue to foster relationships with other research groups and faculty in related research areas (e.g., William Cohen, Alon Lavie, Lori Levin, Tom Mitchell, Carolyn Rosé, and Stephan Vogel).

2.2 Teaching

As a graduate student, I collaboratively taught two short courses with graduate student colleagues. (One of those courses, “Empirical Research Methods in CS,” was motivated by my collaborator's and my sense that something was missing from the core CS curriculum.) I also participated as a lab instructor at the Johns Hopkins annual summer school on language technologies, co-designing and co-leading with my advisor a novel competitive exercise in which students worked in teams to design probabilistic grammars of English. This exercise was described in a paper at a workshop on pedagogy in NLP [60]. I recognized at the time that these unique opportunities helped prepare me for an academic career, and through them I learned hands-on basic lessons about the nuts and bolts of preparing assignments, presenting internally and globally coherent lectures, office hour mentoring, and grading.

At Carnegie Mellon, my teaching mandate is somewhat different from a typical tenure-track faculty member. Because my administrative home is the LTI, which does not have its own undergraduate program, and which has almost no required courses for graduate students, I have been given freedom to teach what I choose.

Undergraduate course. On my own initiative, I designed an undergraduate course in NLP for CS students, giving a broad overview of the field. The course covers most of the most recent edition of the textbook by Jurafsky and Martin, excluding the speech processing sections, since speech is taught in a separate class.

In designing the course, I was careful to avoid pitfalls often made by researchers who teach undergraduates for the first time (including my own earlier mistakes). The weekly assignments are short with no extensions. Weekly readings are short and require a brief response in a two-day window. Pop quizzes encourage regular attendance. Each assignment, reading, and quiz is worth only 1% of the final grade, and they are graded mainly on effort, with an emphasis on discussion and feedback in class. This makes class exercises

more like regular low-stakes workouts than typical “marathon” graduate-style assessments. This system encourages regular attention to the course but de-emphasizes grades, so students can focus on learning. I also distribute a one-page summary at each lecture, so students can *think* and respond during the lecture, rather than rushing to take down notes. Feedback has been positive, the course has done well in attracting students despite heavy competition from other CS electives,² and most students participate actively in class.

Just as I believe concrete applications drive the field of NLP forward, a major *competitive project* that synthesizes the ideas from the lectures drives the educational experience in my undergraduate course. Students work in teams to build (i) automatic question answering systems (without document retrieval; the system is given a document and a natural language question) and (ii) automatic question *asking* systems to stump the other teams’ answerers. At the end of the semester students perform (blind) annotation of question and answer quality, learning hands-on about system evaluation. In 2008 and 2010 I coordinated with Rebecca Hwa at the nearby University of Pittsburgh, and teams from her NLP course joined the competition. The project is described in an NSF workshop paper, [50].

Graduate course and extensions. Every fall since 2006, I have taught an advanced graduate course on statistical NLP, titled “Language and Statistics II.”³ It covers a wide range of statistical models and prediction and learning algorithms for linguistic data. The course includes a unique element: instead of a project, students complete a detailed literature review paper, usually an overview of a current research topic in NLP. This gives graduate students much-needed practice with clear writing, careful reading, and synthesis of disparate technical ideas into coherence.

I recently presented an invited tutorial on linguistic analysis as structured prediction at the International Conference on Machine Learning, condensing the course and focusing its content for the Machine Learning community.

My forthcoming book [12] synthesizes my lectures for this course. The book is drafted at this writing (available on request) and will be sent to the publisher in the summer of 2010. I expect the book to be available electronically and in print in 2010.

In the future I hope to adapt Language and Statistics II into a more general course on structured prediction that will be of interest to machine learning and computational biology students as well as computational linguists.

Graduate seminars. I created a graduate reading seminar in NLP in spring 2009. We read a mix of classical NLP papers and more recent research, covering a wide range of topics. I repeated the offering in spring 2010, this time reading five recent dissertations in NLP, to help students learn what is expected in a Ph.D. thesis.

This semester I created a seminar/project hybrid course on text-driven forecasting. We read papers from CS, Finance, Political Science, and Psychology, with the common theme of using text data to make concrete predictions about the future whose accuracy can be objectively measured. Projects from this course led to two high-profile publications [57, 59].

Looking ahead. I have become experienced and adept at teaching advanced undergraduate and graduate courses in NLP. In fall 2010 I will teach CMU’s course on probabilistic graphical models (in MLD). I am looking forward to this experience as a chance to improve my understanding of the material for use in my ongoing research in structured machine learning.

²This course has been taught three times to classes of 15–25.

³The prerequisite, “Language and Statistics I,” covers the basics of statistical language modeling and discrete machine learning.

3 Service

At CMU, I have served as the organizer of the Intelligence Seminar (an SCS-wide series) for two years. I consistently serve on the LTI graduate admissions committee since my first year at CMU, and have also been active on the LTI curriculum committee since its formation this past year. Apart from my own students, I have served on (or am serving on) nine Ph.D. thesis committees, including one at another university.

I am on the editorial board of the *Computational Linguistics* journal, and frequently review for other top journals (*Artificial Intelligence Journal*, *Journal of Machine Learning Research*, *IEEE Intelligent Systems*, *Proceedings of the National Academy of Sciences*, and others). I have served on numerous NLP and ML conference program committees, often as a senior member of the program committee (“area chair”: EMNLP 2010, NAACL 2010, ACL 2009) or an organization chair (publications chair for ACL 2008, workshop chair for COLING 2010, student travel awards chair for ICML 2008). I have also reviewed regularly for the National Science Foundation.

In 2009 I gave a tutorial at ICML, aiming to bridge the gap between structured machine learning and natural language processing.

When my research leads to reusable datasets, I have consistently made these available to the research community. Examples include forecasting datasets that link text documents to real-world measurements, like film reviews and revenue [61] and financial statements and stock market statistics [62]. Annotation is not a major emphasis of my research, but when new datasets have been gathered and cleaned [63], or small-scale annotation projects completed [64, 65], I have shared them publicly.

NLP research also relies heavily on sharing of implemented tools. I have made available numerous tools for research. Examples from my time before CMU are the Egypt toolkit for statistical machine translation [66]—the Giza word alignment training module survives in most statistical MT systems today—and the Dyna programming language [67]. At CMU, my students and I have released a wide range of open-source NLP tools:

- DAGEEM, a tool for unsupervised dependency grammar induction [68];
- TURBOPARSER [69] and MSTPARSERSTACKED [70], state-of-the-art dependency parsers based, respectively, on integer linear programming and stacking;
- SEMAFOR, a state-of-the-art frame-semantic parser [71]; and
- QUIPU, an experimental statistical machine translation decoder based on quasi-synchronous grammars [72].

The release of software tools and new datasets is important, at the very least to allow other researchers to replicate experiments, but also for application to new problems, and hopefully to allow extensions and improvements on our ideas. Yet it is often difficult to encourage graduate students to devote the resources to this step in the research process. All of my students have publicly released software tools or data, or have plans to do so in the near future.

References

- [1] Noah A. Smith and Jason Eisner. [Annealing techniques for unsupervised statistical language learning](#). In **Proceedings of the Annual Meeting of the Association for Computational Linguistics**, pages 487–494, Barcelona, Spain, July 2004.
- [2] David A. Smith and Noah A. Smith. [Bilingual parsing with factored estimation: Using English to parse Korean](#). In **Proceedings of the Conference on Empirical Methods in Natural Language Processing**, pages 49–56, Barcelona, Spain, July 2004.

- [3] Noah A. Smith and Jason Eisner. [Contrastive estimation: Training log-linear models on unlabeled data](#). In **Proceedings of the Annual Meeting of the Association for Computational Linguistics**, pages 354–362, Ann Arbor, MI, June 2005.
Nominated for best paper award.
- [4] Markus Dreyer, David A. Smith, and Noah A. Smith. [Vine parsing and minimum risk reranking for speed and precision](#). In **Proceedings of the Conference on Natural Language Learning**, pages 201–205, New York, NY, June 2006.
- [5] Noah A. Smith and Jason Eisner. [Annealing structural bias in multilingual weighted grammar induction](#). In **Proceedings of the International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics**, pages 569–576, Sydney, Australia, July 2006.
- [6] Noah A. Smith, David A. Smith, and Roy W. Tromble. [Context-based morphological disambiguation with random fields](#). In **Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing**, pages 475–482, Vancouver, BC, October 2005.
- [7] Yaser Al-Onaizan, Jan Cuřin, Michael Jahr, Kevin Knight, John Lafferty, I. Dan Melamed, Noah A. Smith, Franz-Josef Och, David Purdy, and David Yarowsky. [Statistical machine translation](#). CLSP Research Notes 42, Johns Hopkins University, Baltimore, MD, 1999.
- [8] Noah A. Smith. [From words to corpora: Recognizing translation](#). In **Proceedings of the Conference on Empirical Methods in Natural Language Processing**, pages 95–102, Philadelphia, PA, July 2002.
- [9] Philip Resnik and Noah A. Smith. [The Web as a parallel corpus](#). *Computational Linguistics*, 29(3):349–380, September 2003.
Substantially extends [8].
- [10] Jason Eisner, Eric Goldlust, and Noah A. Smith. [Dyna: A declarative language for implementing dynamic programs](#). In **Proceedings of the Annual Meeting of the Association for Computational Linguistics, companion volume**, pages 218–221, Barcelona, Spain, July 2004.
- [11] Jason Eisner, Eric Goldlust, and Noah A. Smith. [Compiling Comp Ling: Practical weighted dynamic programming and the Dyna language](#). In **Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing**, pages 281–290, Vancouver, BC, October 2005.
- [12] Noah A. Smith. **Linguistic Structure Prediction**. Synthesis Lectures on Human Language Technologies. Morgan and Claypool, to appear.
- [13] Shay B. Cohen, Robert J. Simmons, and Noah A. Smith. [Dynamic programming algorithms as products of weighted logic programs](#). In **Proceedings of the International Conference on Logic Programming**, Udine, Italy, December 2008.
Best student paper award.
- [14] Shay B. Cohen, Robert J. Simmons, and Noah A. Smith. [Products of weighted logic programs](#). *Theory and Practice of Logic Programming*, to appear.
Substantially extends [13].
- [15] Ashish Venugopal, Andreas Zollmann, Noah A. Smith, and Stephan Vogel. [Preference grammars: Softening syntactic constraints to improve statistical machine translation](#). In **Proceedings of the North American Association for Computational Linguistics Human Language Technologies Conference**, pages 236–244, Boulder, CO, May/June 2009.

- [16] Noah A. Smith and Mark Johnson. [Weighted and probabilistic context-free grammars are equally expressive](#). *Computational Linguistics*, 33(4):477–491, December 2007.
- [17] Shay B. Cohen and Noah A. Smith. Empirical risk minimization with data dependent probabilistic grammars. In **Neural Information Processing Systems**, in review.
- [18] Shay B. Cohen and Noah A. Smith. [Joint morphological and syntactic disambiguation](#). In **Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning**, pages 208–217, Prague, Czech Republic, June 2007.
- [19] Kevin Gimpel and Noah A. Smith. [Cube summing, approximate inference with non-local features, and dynamic programming without semirings](#). In **Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics**, pages 157–166, Athens, Greece, March/April 2009.
- [20] André F. T. Martins, Noah A. Smith, and Eric P. Xing. [Concise integer linear programming formulations for dependency parsing](#). In **Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing**, pages 342–350, Singapore, August 2009. **Best paper award.**
- [21] David A. Smith and Noah A. Smith. [Probabilistic models of nonprojective dependency trees](#). In **Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning**, pages 132–140, Prague, Czech Republic, June 2007.
- [22] Jason Eisner and Noah A. Smith. [Favor short dependencies: Parsing with soft and hard constraints on dependency length](#). In Harry Bunt, Paola Merlo, and Joakim Nivre, editors, **Trends in Parsing Technology**, Text, Speech, and Language Technology, chapter 9. Springer, to appear. Originally published as “Parsing with soft and hard constraints on dependency length” in **Proceedings of the International Workshop on Parsing Technologies**, pages 30–41, Vancouver, BC, October 2005.
- [23] Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. [Probabilistic frame-semantic parsing](#). In **Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference**, Los Angeles, CA, June 2010.
- [24] Desai Chen, Nathan Schneider, Dipanjan Das, and Noah A. Smith. [SEMAFOR: Frame argument resolution with log-linear models](#). In **Proceedings of the International (ACL) Workshop on Semantic Evaluations**, Uppsala, Sweden, July 2010.
- [25] Michael Heilman and Noah A. Smith. [Tree edit models for recognizing textual entailments, paraphrases, and answers to questions](#). In **Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference**, Los Angeles, CA, June 2010.
- [26] Dipanjan Das and Noah A. Smith. [Paraphrase identification as probabilistic quasi-synchronous recognition](#). In **Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing**, pages 468–476, Singapore, August 2009.
- [27] André F. T. Martins, Mário A. T. Figueiredo, Pedro M. Q. Aguiar, Noah A. Smith, and Eric P. Xing. [Nonextensive entropic kernels](#). In **Proceedings of the International Conference on Machine Learning**, pages 640–647, Helsinki, Finland, July 2008.
- [28] André F. T. Martins, Noah A. Smith, Eric P. Xing, Mário A. T. Figueiredo, and Pedro M. Q. Aguiar. [Nonextensive information theoretic kernels on measures](#). *Journal of Machine Learning Research*,

10:935–975, April 2009.
Substantially extends [27].

- [29] André F. T. Martins, Noah A. Smith, Eric P. Xing, Pedro M. Q. Aguiar, and Mário A. T. Figueiredo. Online multiple kernel learning. In **Neural Information Processing Systems**, in review.
- [30] Noah A. Smith. *Novel Estimation Methods for Unsupervised Discovery of Latent Structure in Natural Language Text*. Ph.D. thesis, Department of Computer Science, Johns Hopkins University, Baltimore, MD, October 2006. Supervised by Jason Eisner.
- [31] André F. T. Martins, Noah A. Smith, and Eric P. Xing. [Polyhedral outer approximations with application to natural language parsing](#). In **Proceedings of the International Conference on Machine Learning**, pages 713–720, Montréal, Québec, June 2009.
- [32] André F. T. Martins, Dipanjan Das, Noah A. Smith, and Eric P. Xing. [Stacking dependency parsers](#). In **Proceedings of the Conference on Empirical Methods in Natural Language Processing**, pages 157–166, Waikiki, HI, October 2008.
- [33] Noah A. Smith, Douglas L. Vail, and John D. Lafferty. [Computationally efficient M-estimation of log-linear structure models](#). In **Proceedings of the Annual Meeting of the Association for Computational Linguistics**, pages 752–759, Prague, Czech Republic, June 2007.
- [34] Kevin Gimpel, Dipanjan Das, and Noah A. Smith. [Distributed asynchronous online learning for natural language processing](#). In **Proceedings of the Conference on Computational Natural Language Learning**, Uppsala, Sweden, July 2010.
- [35] Kevin Gimpel and Noah A. Smith. [Softmax-margin CRFs: Training log-linear models with loss functions](#). In **Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference**, Los Angeles, CA, June 2010.
- [36] André F. T. Martins, Kevin Gimpel, Noah A. Smith, Eric P. Xing, Pedro M. Q. Aguiar, and Mário A. T. Figueiredo. Aggressive online learning of structured classifiers. In **Neural Information Processing Systems**, in review.
- [37] Shay B. Cohen, Kevin Gimpel, and Noah A. Smith. [Logistic normal priors for unsupervised probabilistic grammar induction](#). In **Advances in Neural Information Processing Systems 21**, pages 321–328, Vancouver, BC, December 2008.
- [38] Shay B. Cohen and Noah A. Smith. [Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction](#). In **Proceedings of the North American Association for Computational Linguistics Human Language Technologies Conference**, pages 74–82, Boulder, CO, May/June 2009.
- [39] Shay B. Cohen and Noah A. Smith. [Variational inference for grammar induction with prior knowledge](#). In **Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing, companion volume**, pages 1–4, Singapore, August 2009.
- [40] Shay B. Cohen and Noah A. Smith. [Covariance in unsupervised learning of probabilistic grammars](#). *Journal of Machine Learning Research*, in review.
Substantially extends [38].
- [41] Shay B. Cohen, David M. Blei, and Noah A. Smith. [Variational inference for adaptor grammars](#). In **Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference**, Los Angeles, CA, June 2010.

- [42] ThuyLinh Nguyen, Stephan Vogel, and Noah A. Smith. Nonparametric word segmentation for machine translation. In **Proceedings of the International Conference on Computational Linguistics**, Beijing, China, August 2010.
- [43] Shay B. Cohen and Noah A. Smith. [Viterbi training for PCFGs: Hardness results and competitiveness of uniform initialization](#). In **Proceedings of the Annual Meeting of the Association for Computational Linguistics**, Uppsala, Sweden, July 2010.
- [44] Kevin Gimpel and Noah A. Smith. [Rich source-side context for statistical machine translation](#). In **Proceedings of the ACL Workshop on Statistical Machine Translation**, pages 9–17, Columbus, OH, June 2008.
- [45] Ashish Venugopal, Andreas Zollmann, Noah A. Smith, and Stephan Vogel. [Wider pipelines: \$N\$ -best alignments and parses in MT training](#). In **Proceedings of the Conference of the Association for Machine Translation in the Americas**, Waikiki, HI, October 2008.
- [46] Kevin Gimpel and Noah A. Smith. [Feature-rich translation by quasi-synchronous lattice parsing](#). In **Proceedings of the Conference on Empirical Methods in Natural Language Processing**, pages 219–228, Singapore, August 2009.
- [47] Sourish Chaudhuri, Naman K. Gupta, Noah A. Smith, and Carolyn P. Rosé. [Leveraging structural relations for fluent compressions at multiple compression rates](#). In **Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing, companion volume**, pages 101–104, Singapore, August 2009.
- [48] André F. T. Martins and Noah A. Smith. [Summarization with a joint model for sentence extraction and compression](#). In **Proceedings of the NAACL-HLT Workshop on Integer Linear Programming for Natural Language Processing**, Boulder, CO, June 2009.
- [49] Mengqiu Wang, Noah A. Smith, and Teruko Mitamura. [What is the Jeopardy model? a quasi-synchronous grammar for QA](#). In **Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning**, pages 22–32, Prague, Czech Republic, June 2007.
Nominated for best paper award.
- [50] Noah A. Smith, Michael Heilman, and Rebecca Hwa. [Question generation as a competitive undergraduate course project](#). In **Proceedings of the NSF Workshop on the Question Generation Shared Task and Evaluation Challenge**, Arlington, VA, September 2008.
- [51] Michael Heilman and Noah A. Smith. [Extracting simplified statements for factual question generation](#). In **Proceedings of the AIED Workshop on Question Generation**, Pittsburgh, PA, June 2010.
- [52] Michael Heilman and Noah A. Smith. [Good question! statistical ranking for question generation](#). In **Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference**, Los Angeles, CA, June 2010.
- [53] Michael Heilman and Noah A. Smith. [Ranking automatically generated questions as a shared task](#). In **Proceedings of the AIED Workshop on Question Generation**, Brighton, UK, July 2009.
- [54] Tae Yano, William W. Cohen, and Noah A. Smith. [Predicting response to political blog posts with topic models](#). In **Proceedings of the North American Association for Computational Linguistics Human Language Technologies Conference**, pages 477–485, Boulder, CO, May/June 2009.
- [55] Tae Yano and Noah A. Smith. [What’s worthy of comment? content and comment volume in political blogs](#). In **Proceedings of the International AAAI Conference on Weblogs and Social Media**, Washington, DC, May 2010.

- [56] Tae Yano, Philip Resnik, and Noah A. Smith. [Shedding \(a thousand points of\) light on biased language](#). In **Proceedings of the NAACL-HLT Workshop on Creating Speech and Language Data With Mechanical Turk**, Los Angeles, CA, June 2010.
- [57] Brendan O’Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. [From tweets to polls: Linking text sentiment to public opinion time series](#). In **Proceedings of the International AAAI Conference on Weblogs and Social Media**, Washington, DC, May 2010.
- [58] Shimon Kogan, Dmitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. [Predicting risk from financial reports with regression](#). In **Proceedings of the North American Association for Computational Linguistics Human Language Technologies Conference**, pages 272–280, Boulder, CO, May/June 2009.
- [59] Mahesh Joshi, Dipanjan Das, Kevin Gimpel, and Noah A. Smith. [Movie reviews and revenues: An experiment in text regression](#). In **Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference**, Los Angeles, CA, June 2010.
- [60] Jason Eisner and Noah A. Smith. [Competitive grammar writing](#). In **Proceedings of the ACL Workshop on Issues in Teaching Computational Linguistics**, pages 97–105, Columbus, OH, June 2008.
- [61] [MOVIE\\$ CORPUS](#), developed by Mahesh Joshi, Dipanjan Das, and Kevin Gimpel. Collection of pre-release movie reviews, metadata, and opening weekend revenues, 2010, see [59].
- [62] [10-K CORPUS](#), developed with three others. Collection of 10-K reports and preceding and following stock return volatility measurements, 2009, see [58].
- [63] [POLITICAL BLOG CORPUS](#), developed by Tae Yano. Text collection from five American political blogs, 2009, see [54].
- [64] [CURD](#), developed by Dan Tasse. Corpus of semantically annotated recipes, 2008, see [73].
- [65] [AMAZON MECHANICAL TURK POLITICAL BIAS DATA](#), developed by Tae Yano. Sentences from political blogs with crowdsourced annotations of political bias, 2010, see [56].
- [66] [EGYPT](#), developed with nine others. Toolkit for statistical machine translation, including GIZA training module and CAIRO word alignment visualizer, 1999, see [7, 74].
- [67] [DYNA](#), developed with five others. Declarative programming language for weighted dynamic programming, 2004, see [10, 11].
- [68] [DAGEEM](#), developed by Shay Cohen. Unsupervised dependency grammar induction, 2008, see [37].
- [69] [TURBOPARSER](#), developed by André Martins. Multilingual dependency parser, 2009, see [20].
- [70] [MSTPARSER](#), [STACKED](#), developed by André Martins and Dipanjan Das. Multilingual dependency parser, 2008, see [32].
- [71] [SEMAFOR](#), developed by Dipanjan Das, Nathan Schneider, and Desai Chen. Frame-semantic parser, 2010, see [23].
- [72] [QUIPU](#), developed by Kevin Gimpel. Statistical machine translation system, 2009, see [46].
- [73] Dan Tasse and Noah A. Smith. [SOUR CREAM: Toward semantic processing of recipes](#). Technical Report CMU-LTI-08-005, Carnegie Mellon University, Pittsburgh, PA, May 2008.
- [74] Noah A. Smith and Michael E. Jahr. [Cairo: An alignment visualization tool](#). In **Proceedings of the Language Resources and Evaluation Conference**, pages 549–552, Athens, Greece, May/June 2000.