Natural Language Processing for the Long Tail*

David Bamman
School of Information
University of California, Berkeley
dbamman@berkeley.edu

Natural language processing (NLP) is a research area that stands at the intersection of linguistics and computer science; its focus is the development of automatic methods that can reason about the internal structure of text. This includes **part-of-speech tagging** (which, for a sentence like *John ate the apple*, infers that *John* is a noun, and *ate* a verb), **syntactic parsing** (which infers that *John* is the syntactic subject of *ate*, and *the apple* its direct object), and **named entity recognition** (which infers that *John* is a PERSON, and that *apple* is not, for example, an ORGANIZATION of the same name). Beyond these core tasks, NLP also encompasses sentiment analysis, named entity linking, information extraction, and machine translation (among many other applications).

Over the past few years, NLP has become an increasingly important element in computational research in the humanities. Automatic part-of-speech taggers have been used to filter input in topic models (Jockers, 2013) and explore poetic enjambment (Houston, 2014). Syntactic parsers have been used to help select relevant context for concordances (Benner, 2014). Named entity recognizers have been used to map the attention given to various cities in American fiction (Wilkens, 2013) and to map toponyms in Joyce's *Ulysses* (Derven et al., 2014) and Pelagios texts (Simon et al., 2014). The sequence tagging models behind many part-of-speech taggers have also been used for identifying genres in books (Underwood et al., 2013).

There is a substantial gap, however, between the quality of the NLP used by researchers in the humanities and the state of the art. Research in natural language processing has overwhelmingly focused much of its attention on English, and specifically on the domain of news (simply as a function of the availability of training data). The Penn Treebank (Marcus et al., 1993)—containing morphosyntactic annotations of the *Wall Street Journal*—has driven automatic parsing performance in English above 92% (Andor et al., 2016); part-of-speech tagging on this same data now yields accuracies over 97% (Søgaard, 2011). While a handful of other high-resource languages (German, French, Spanish, Japanese) have attained comparable performance on similar data (Hajič et al., 2009), many languages simply have too few resources (or none whatsoever) to train robust automatic tools. Even within English, out-of-domain performance of many NLP tasks—in which, for example, a syntactic parser trained on the *Wall Street Journal* is used to automatically label the syntax for *Paradise Lost*—is bleak. Figure 1 illustrates one sentence from *Paradise Lost* automatically tagged and parsed using a tool trained on the *Wall Street Journal*. Since this model is trained on newswire, it expects newswire as its input; errors in the part-of-speech assignment snowball to bigger errors in syntax.

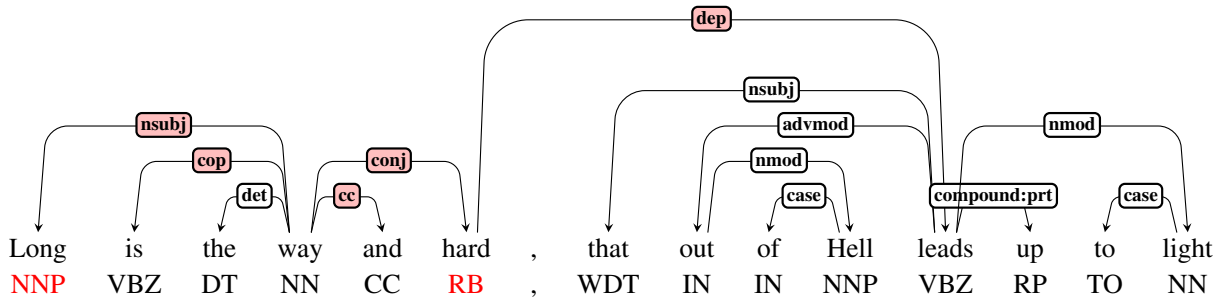

Figure 1: Parsers and part-of-speech taggers trained on the *WSJ* expect newswire syntax. Automatically parsed syntactic dependency graph with part-of-speech tags for *Long is the way and hard, that out of Hell leads up to light*. Errors in part-of-speech tags and dependency arcs are shown in red. Part-of-speech errors snowball into major syntactic errors.

*Paper presented at *Digital Humanities* 2017.

Table 1 provides a summary of recent research that has investigated the disparity between training data and test data for several NLP tasks (including part-of-speech tagging, syntactic parsing and named entity recognition). While many of these tools are trained on the same fixed corpora (comprised primarily of newswire), they suffer a dramatic drop in performance when used to analyze texts that come from a substantially different domain. Without any form of adaptation (such as normalizing spelling across time spans), the performance of an out-of-the-box part-of-speech tagger can, at worse, be half that of its performance on contemporary newswire. On average, differences in style amount to a drop in performance of approximately 10-20 absolute percentage points across tasks. These are substantial losses.

Citation	Task	In domain	Acc.	Out domain	Acc.
Rayson et al. (2007)	POS	English news	97.0%	Shakespeare	81.9%
Scheible et al. (2011)	POS	German news	97.0%	Early Modern German	69.6%
Moon and Baldridge (2007)	POS	WSJ	97.3%	Middle English	56.2%
Pennacchiotti and Zanzotto (2008)	POS	Italian news	97.0%	Dante	75.0%
Derczynski et al. (2013b)	POS	WSJ	97.3%	Twitter	73.7%
Yang and Eisenstein (2016)	POS	WSJ		Early Modern English	74.3%
Gildea (2001)	PS parsing	WSJ	86.3 F	Brown corpus	80.6 F
Lease and Charniak (2005)	PS parsing	WSJ	89.5 F	GENIA medical texts	76.3 F
Burga et al. (2013)	Dep. parsing	WSJ	88.2%	Patent data	79.6%
Pekar et al. (2014)	Dep. parsing	WSJ	86.9%	Broadcast news	79.4%
				Magazines	77.1%
				Broadcast conversation	73.4%
Derczynski et al. (2013a)	NER	CoNLL 2003	89.0 F	Twitter	41.0 F

Table 1: Out-of-domain performance for several NLP tasks, including POS tagging, phrase structure (PS) parsing, dependency parsing and named entity recognition. Accuracies are reported in percentages; phrase structure parsing and NER are reported in F1 measure.

While many techniques are currently under development in the NLP community for domain adaptation (Blitzer et al., 2006; Chelba and Acero, 2006; Daumé III, 2009; Glorot et al., 2011; Yang and Eisenstein, 2014), including leveraging fortuitous data (Plank, 2016), they often require specialized expertise that can be a bottleneck for researchers in the humanities. The simplest and most empowering solution is often to create *in-domain* data and train NLP methods on it directly; in-domain data can substantially increase performance, almost to levels approaching state-of-the-art on newswire (Scheible et al., 2011; Moon and Baldridge, 2007; Derczynski et al., 2013b; Strötgen and Gertz, 2012).

When adding training data of Early Modern German and adding spelling normalization, Scheible et al. (2011) increase POS tagging accuracy on Early Modern German texts from 69.6% to 91.0%; when Moon and Baldridge (2007) train a POS tagger on Middle English texts, this pushes their accuracy from 56.2% to 93.7%; when Derczynski et al. (2013b) train a POS tagger directly on Twitter data, this increases accuracy from 73.7% to 88.4%. In-domain data is astoundingly helpful for many NLP tasks, from part-of-speech tagging and syntactic parsing to temporal tagging (Strötgen and Gertz, 2012).

The difficulty, of course, is that training data is expensive to create at scale since it relies on human judgments; and the cost of this data scales with the complexity of the task, so that morphosyntactic or semantic annotations (which require a holistic understanding of an entire sentence) are often prohibitive. Few projects achieve this scale for domains in the humanities, but when they do, they have real impact – these include WordHoard, which contains part-of-speech annotations for Shakespeare, Chaucer and Spenser (Mueller, 2015); the Penn and York parsed corpora of historical English (Taylor and Kroch, 2000; Kroch et al., 2004; Taylor et al., 2006); the Perseus Greek and Latin treebanks (Bamman and Crane, 2011), which contain morphosyntactic annotations for classical Greek and Latin works; the Index Thomisticus (Passarotti, 2007), which contains morphosyntactic annotations for the works of Thomas Aquinas; the PROIEL treebank (Haug and Jøhndal, 2008), which contains similar annotations for several translations of the Bible (Greek, Latin, Gothic, Armenian and Church Slavonic); the Tycho Brahe Parsed Corpus of Historical Portuguese (Galves and Faria, 2010); the Icelandic Parsed Historical Corpus (Rögnvaldsson et al., 2012), and Twitter, annotated for part-of-speech (Gimpel et al., 2011) and dependency syntax (Kong et al., 2014).

The availability of these annotated corpora means that we have the ability to train NLP tools for some dialects, domains and genres of Ancient Greek, Latin, Early Modern English, historical Portuguese, and a few other languages; this doesn't help the scholar working on John Milton, Virginia Woolf, Miguel Cervantes, or the countless other authors

and genres in the long tail of underserved domains that researchers are increasingly finding high-quality NLP useful to help analyze. In this talk, I"ll argue for an alternative: an open repository of linguistic annotations that scholars can use to train statistical models for processing natural language in a variety of domains, leveraging information from complementary sources (such as the works of Shakespeare) to perform well on a target domain of interest (such as the works of Christopher Marlowe). What this repository critically relies on is the expertise of the individuals who simultaneously are the consumers of NLP for their long-tail domain and are in the uniquely best position to create linguistic data to support their own work—and in doing so, can help develop an ecosystem that can support the work of others.

Acknowledgments

Many thanks to the anonymous reviewers for helpful feedback. This work is supported by a grant by the Digital Humanities at Berkeley initiative.

References

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. Globally normalized transition-based neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2442–2452, Berlin, Germany, August 2016. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P16-1231`.

David Bamman and Gregory Crane. The ancient Greek and Latin dependency treebanks. In *Language Technology for Cultural Heritage*, pages 79–98. Springer, 2011.

Drayton Callen Benner. Marrying the benefits of print and digital: Algorithmically selecting context for a key word. In *Digital Humanities 2014*, 2014.

John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 120–128, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. ISBN 1-932432-73-6. URL `http://dl.acm.org/citation.cfm?id=1610075.1610094`.

Alicia Burga, Joan Codina, Gabriella Ferraro, Horacio Saggion, and Leo Wanner. The challenge of syntactic dependency parsing adaptation for the patent domain. In *ESSLLI-13 Workshop on Extrinsic Parse Improvement*, 2013.

Ciprian Chelba and Alex Acero. Adaptation of maximum entropy capitalizer: Little data can help a lot. *Computer Speech & Language*, 20(4):382–399, 2006.

Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

Leon Derczynski, Diana Maynard, Niraj Aswani, and Kalina Bontcheva. Microblog-genre noise and impact on semantic annotation accuracy. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 21–30. ACM, 2013a.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *RANLP*, pages 198–206, 2013b.

Caleb Derven, Aja Teehan, and John Keating. Mapping and unmapping Joyce: Geoparsing wandering rocks. In *Digital Humanities 2014*, 2014.

Charlotte Galves and Pablo Faria. Tycho Brahe Parsed Corpus of Historical Portuguese. http://www.tycho.iel.unicamp.br/corpus/en/index.html, 2010.

Daniel Gildea. Corpus variation and parser performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 167–202, 2001.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, companion volume*, Portland, OR, June 2011. URL `http://www.cs.cmu.edu/~nasmith/papers/gimpel+etal.acl11.pdf`.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 513–520, 2011.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, et al. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–18. Association for Computational Linguistics, 2009.

Dag TT Haug and Marius Jøhndal. Creating a parallel treebank of the old indo-european bible translations. In *Proceedings of the Language Technology for Cultural Heritage Data Workshop (LaTeCH 2008), Marrakech, Morocco, 1st June 2008*, pages 27–34, 2008.

Natalie Houston. Enjambment and the poetic line: Towards a computational poetics. In *Digital Humanities 2014*, 2014.

Matthew Jockers. "Secret" recipe for topic modeling themes. http://www.matthewjockers.net/2013/04/12/secret-recipe-for-topic-modeling-themes/, April 2013.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012, Doha, Qatar, October 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D14-1108`.

Anthony Kroch, Beatrice Santorini, and Lauren Delfs. Penn-Helsinki parsed corpus of Early Modern English. *Philadelphia: Department of Linguistics, University of Pennsylvania*, 2004.

Matthew Lease and Eugene Charniak. Parsing biomedical literature. In *Natural Language Processing–IJCNLP 2005*, pages 58–69. Springer, 2005.

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

Taesun Moon and Jason Baldridge. Part-of-speech tagging for Middle English through alignment and projection of parallel diachronic texts. In *EMNLP-CoNLL*, pages 390–399, 2007.

Martin Mueller. Wordhoard. http://wordhoard.northwestern.edu/, Accessed 2015.

Marco Passarotti. Verso il Lessico Tomistico Biculturale. La treebank dell'Index Thomisticus. In Petrilli Raffaella and Femia Diego, editors, *Il filo del discorso. Intrecci testuali, articolazioni linguistiche, composizioni logiche. Atti del XIII Congresso Nazionale della Società di Filosofia del Linguaggio, Viterbo, Settembre 2006*, pages 187–205. Roma, Aracne Editrice, Pubblicazioni della Società di Filosofia del Linguaggio, 2007.

Viktor Pekar, Juntao Yu, Mohab El-karef, and Bernd Bohnet. Exploring options for fast domain adaptation of dependency parsers. *SPMRL-SANCL 2014*, page 54, 2014.

Marco Pennacchiotti and Fabio Massimo Zanzotto. Natural language processing across time: An empirical investigation on italian. In *Advances in natural language processing*, pages 371–382. Springer, 2008.

Barbara Plank. What to do about non-standard (or non-canonical) language in NLP. In *KONVENZ*, 2016. URL `http://arxiv.org/abs/1608.07836`.

Paul Rayson, Dawn Archer, Alistair Baron, Jonathan Culpeper, and Nicholas Smith. Tagging the bard: Evaluating the accuracy of a modern pos tagger on early modern english corpora. In *Proceedings of Corpus Linguistics (CL2007)*, 2007.

Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson, and Joel Wallenberg. The Icelandic Parsed Historical Corpus (IcePaHC). In *LREC*, pages 1977–1984, 2012.

Silke Scheible, Richard J Whitt, Martin Durrell, and Paul Bennett. Evaluating an 'off-the-shelf' POS-tagger on early modern German text. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*, pages 19–23. Association for Computational Linguistics, 2011.

Rainer Simon, Elton T. E. Barker, Pau de Soto, and Leif Isaksen. Pelagios 3: Towards the semi- automatic annotation of toponyms in early geospatial documents. In *Digital Humanities 2014*, 2014.

Anders Søgaard. Semisupervised condensed nearest neighbor for part-of-speech tagging. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 48–52. Association for Computational Linguistics, 2011.

Jannik Strötgen and Michael Gertz. Temporal tagging on different domains: Challenges, strategies, and gold standards. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Ann Taylor and Anthony S Kroch. The Penn-Helsinki Parsed Corpus of Middle English. *University of Pennsylvania*, 2000.

Ann Taylor, Arja Nurmi, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen. Parsed Corpus of Early English Correspondence. Oxford Text Archive, 2006.

Ted Underwood, Michael L Black, Loretta Auvil, and Boris Capitanu. Mapping mutable genres in structurally complex volumes. In *Big Data, 2013 IEEE International Conference on*, pages 95–103. IEEE, 2013.

Matthew Wilkens. The geographic imagination of Civil War-era American fiction. *American Literary History*, 25 (4):803–840, 2013. doi: 10.1093/alh/ajt045. URL `http://alh.oxfordjournals.org/content/25/4/803.short`.

Yi Yang and Jacob Eisenstein. Fast easy unsupervised domain adaptation with marginalized structured dropout. *Proceedings of the Association for Computational Linguistics (ACL), Baltimore, MD*, 2014.

Yi Yang and Jacob Eisenstein. Part-of-speech tagging for historical english. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1328, San Diego, California, June 2016. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N16-1157`.