A Universal Part-of-Speech Tagset

Slav Petrov[1] **Dipanjan Das**[2] **Ryan McDonald**[1]

[1]Google Research, New York, NY, USA, {slav, ryanmcd}@google.com [2]Carnegie Mellon University, Pittsburgh, PA, USA, dipanjan@cs.cmu.edu

Abstract

To facilitate future research in unsupervised induction of syntactic structure and to standardize best-practices, we propose a tagset that consists of twelve universal part-of-speech categories. In addition to the tagset, we develop a mapping from 25 different treebank tagsets to this universal set. As a result, when combined with the original treebank data, this universal tagset and mapping produce a dataset consisting of common parts-of-speech for 22 different languages. We highlight the use of this resource via three experiments, that (1) compare tagging accuracies across languages, (2) present an unsupervised grammar induction approach that does not use gold standard part-of-speech tags, and (3) use the universal tags to transfer dependency parsers between languages, achieving state-of-the-art results.

Keywords: Part-of-Speech Tagging, Multilinguality, Annotation Guidelines

1. Introduction

Part-of-speech (POS) tagging has received a great deal of attention as it is a critical component of most natural language processing systems. As supervised POS tagging accuracies for English (measured on the PennTreebank (Marcus et al., 1993)) have converged to around 97.3% (Toutanova et al., 2003; Shen et al., 2007; Manning, 2011), the attention has shifted to unsupervised approaches (Christodoulopoulos et al., 2010). In particular, there has been growing interest in both multi-lingual POS induction (Snyder et al., 2009; Naseem et al., 2009) and cross-lingual POS induction via projections (Yarowsky and Ngai, 2001; Xi and Hwa, 2005; Das and Petrov, 2011).

Underlying these studies is the idea that a set of (coarse) syntactic POS categories exists in a similar form across languages. These categories are often called *universals* to represent their cross-lingual nature (Carnie, 2002; Newmeyer, 2005). For example, Naseem et al. (2009) use the Multext-East (Erjavec, 2004) corpus to evaluate their multi-lingual POS induction system, because it uses the same tagset for multiple languages. When corpora with common tagsets are unavailable, a standard approach is to manually define a mapping from language and treebank specific fine-grained tagsets to a predefined universal set. This is the approach taken by Das and Petrov (2011) to evaluate their cross-lingual POS projection system.

To facilitate future research and to standardize best-practices, we propose a tagset that consists of twelve universal POS categories. While there might be some controversy about what the exact tagset should be, we feel that these twelve categories cover the most frequent part-of-speech that exist in most languages. In addition to the tagset, we also develop a mapping from fine-grained POS tags for 25 different treebanks to this universal set. As a result, when combined with the original treebank data, this universal tagset and mapping produce a dataset consisting of common parts-of-speech for 22 different languages.[1] Both the tagset and mappings are made available for download at **http://code.google.com/p/universal-pos-tags/**.

This resource serves multiple purposes. First, as mentioned previously, it is useful for building and evaluating unsupervised and cross-lingual taggers and parsers. Second, it permits for a better comparison of accuracy across languages for supervised taggers. Statements of the form "POS tagging for language X is harder than for language Y" are vacuous when the tagsets used for the two languages are incomparable (not to mention of different cardinality). Fi-

[1] We include mappings for two different Chinese, German and Japanese treebanks.

nally, it also permits language technology practitioners to train POS taggers with common tagsets across multiple languages. This in turn facilitates downstream application development as there is no need to maintain language specific rules or systems due to differences in treebank annotation guidelines.

In this paper, we specifically highlight three use cases of this resource. First, using our universal tagset and mapping, we run an experiment comparing POS tagging accuracies for 25 different treebanks on a single tagset. Second, we combine the cross-lingual projection part-of-speech taggers of Das and Petrov (2011) with the grammar induction system of Naseem et al. (2010) – which requires a universal tagset – to produce a completely unsupervised grammar induction system for multiple languages, that does not require gold POS tags or any other type of manual annotation in the target language. Finally, we show that a delexicalized English parser, whose predictions rely solely on the universal POS tags of the input sentence, can be used to parse a foreign language POS sequence, achieving higher accuracies than state-of-the-art unsupervised parsers. These experiments highlight that our universal tagset captures a substantial amount of information and carries that information over across languages boundaries.

2. Tagset

While there might be some disagreement about the exact definition of an universal POS tagset (Evans and Levinson, 2009), several scholars have argued that a set of coarse POS categories (or syntactic universals) exists across languages in one form or another (Carnie, 2002; Newmeyer, 2005). Rather than attempting to define an 'a priori' or 'inherent'

sentence: The oboist Heinz Holliger has taken a hard line about the problems . original: DT NN NNP NNP VBZ VBN DT JJ NN IN DT NNS . universal: DET NOUN NOUN NOUN VERB VERB DET ADJ NOUN ADP DET NOUN . Figure 1: Example English sentence with its language specific and corresponding universal POS tags.

tagset, we took a pragmatic approach during the design of the universal POS tagset and focused our attention on the POS categories that we expect to be most useful (and necessary) for users of POS taggers. In our opinion, these are NLP practitioners using taggers in downstream applications, and NLP researchers using POS taggers in grammar induction and other experiments.

A high-level analysis of the tagsets underlying various treebanks shows that the majority of tagsets are very fine-grained and language specific. This observation has of course been made many times in the past: Smith and Eisner (2005) defined a collapsed set of 17 English POS tags (instead of the 45 tags in the PennTreebank) that has subsequently been adopted by most unsupervised English POS induction work. The organizers of the CoNLL shared tasks on dependency parsing provided coarse (but still language specific) tags in addition to the original fine-grained tags (Buchholz and Marsi, 2006; Nivre et al., 2007). A number of different authors have investigated reduced tagsets that improve tagging and parsing accuracies (Brants, 1995; Dienes and Oravecz, 2000; Dominguez and Infante-Lopez, 2008). Rambow et al. (2006) defined a multilingual tagset that is close to ours and McDonald and Nivre (2007) identified eight different coarse POS tags when analyzing the errors of two dependency parsers across the 13 different languages from the CoNLL shared tasks. Finally, Dickinson and Jochim (2008) investigated methods for comparing tagsets and Zeman (2008) provided a tool for converting between tagsets.

Our universal POS tagset unifies this previous work and extends it to 22 languages, defining the following twelve POS tags: NOUN (nouns), VERB (verbs), ADJ (adjectives), ADV (adverbs), PRON (pronouns), DET (determiners and articles), ADP (prepositions and postpositions), NUM (numerals), CONJ (conjunctions), PRT (particles), '.' (punctuation marks) and X (a catch-all for other categories such as abbreviations or foreign words).

We did not rely on intrinsic definitions of the above categories. Instead, each category is defined operationally. For each treebank under consideration, we studied the exact POS tag definitions and annotation guidelines and created a mapping from the original treebank tagset to these universal POS tags. Most of the decisions were fairly clear. For example, from the PennTreebank, VB, VBD, VBG, VBN, VBP, VBZ and MD (modal) were all mapped to VERB. A less clear case was the universal tag for particles, PRT, which was mapped from POS (possessive), RP (particle) and TO (the word 'to'). In particular, the TO tag is ambiguous in the PennTreebank between infinitival markers and the preposition 'to'. Thus, no automatic mapping can differentiate between the two and as a result some prepositions will be marked as particles in the universal tagset.

Another case we had to consider is that some tag categories do not occur in all languages, or are not explicitly labeled in the treebanks. While all languages have a way of describing the properties of objects (which themselves are typically referred to with nouns), many have argued that Korean does not technically have adjectives, but instead expresses properties of nouns via stative verbs (Kim, 2002). As a result, in our mapping for Korean, we mapped stative verbs to the universal ADJ tag. In other cases this was clearer, e.g. the Bulgarian treebank has no category for determiners or articles. This is not to say that there are no determiners in the Bulgarian language, however, since they are not annotated as such in the treebank, we are not able to include them in our mapping.

Figure 1 gives an example mapping for an English sentence from the PennTreebank. While one might be worried that the universal POS tags are too coarse for downstream applications, at least for dependency parsing this seems not to be the case. A supervised state-of-the-art English dependency parser looses only about 0.6% in accuracy when provided with the 12 universal POS tags instead of the original 45 PennTreebank tags.

In Table 3 at the end of this paper we provide a list of the treebanks that we studied, as well as the actual mappings that we constructed. For space reasons the mappings for treebanks with very large tagsets had to be omitted. Already a quick glance at the table shows that the language-specific tagsets vary in their specificity in different areas. Some tagsets define only a single pronoun category, while others distinguish between a dozen different pronouns. Similarly, many treebanks specify a dozen multiple fine-grained verb categories, while others have a single category. Often times this is not because the language does not exhibit variations in those areas of its grammar, but because the linguists defining the annotation standards for the treebanks choose different trade-offs. Our universal tagset aims to simplify the tags and unify them across languages.

Since its release in the early 2011, the tagset has been used in a number of ways. Das and Petrov (2011) presented a part-of-speech projection system that uses the tagset for evaluating projected POS taggers and Gimpel et al. (2011) used it as the basis of a Twitter annotation project. McDonald et al. (2011) and Cohen et al. (2011) built multilingual parser projection systems that rely on the universal part-of-speech tags for transferring information between languages. Despite the coarseness of the universal tagset, their projected parsers significantly outperformed previous work, highlighting the utility of the universal tagset. We replicate some of the experiments of McDonald et al. (2011) in the next section. Finally, DeNero and Uszkoreit (2011) presented a bilingual grammar induction system for machine translation reordering that uses the universal tags to connect the two languages. Without the universal POS tags, their system suffers significant performance drops.

The tagset mappings are hosted as an open source project at: http://code.google.com/p/universal-pos-tags/. One main

Language	Source	# Tags	O/O	U/U	O/U
Arabic	PADT/CoNLL07 (Hajič et al., 2004)	21	96.1	96.9	97.0
Basque	Basque3LB/CoNLL07 (Aduriz et al., 2003)	64	89.3	93.7	93.7
Bulgarian	BTB/CoNLL06 (Simov et al., 2002)	54	95.7	97.5	97.8
Catalan	CESS-ECE/CoNLL07 (Martí et al., 2007)	54	98.5	98.2	98.8
Chinese	Penn ChineseTreebank 6.0 (Palmer et al., 2007)	34	91.7	93.4	94.1
Chinese	Sinica/CoNLL07 (Chen et al., 2003)	294	87.5	91.8	92.6
Czech	PDT/CoNLL07 (Böhmová et al., 2003)	63	99.1	99.1	99.1
Danish	DDT/CoNLL06 (Kromann et al., 2003)	25	96.2	96.4	96.9
Dutch	Alpino/CoNLL06 (Van der Beek et al., 2002)	12	93.0	95.0	95.0
English	PennTreebank (Marcus et al., 1993)	45	96.7	96.8	97.7
French	FrenchTreebank (Abeillé et al., 2003)	30	96.6	96.7	97.3
German	Tiger/CoNLL06 (Brants et al., 2002)	54	97.9	98.1	98.8
German	Negra (Skut et al., 1997)	54	96.9	97.9	98.6
Greek	GDT/CoNLL07 (Prokopidis et al., 2005)	38	97.2	97.5	97.8
Hungarian	Szeged/CoNLL07 (Csendes et al., 2005)	43	94.5	95.6	95.8
Italian	ISST/CoNLL07 (Montemagni et al., 2003)	28	94.9	95.8	95.8
Japanese	Verbmobil/CoNLL06 (Kawata and Bartels, 2000)	80	98.3	98.0	99.1
Japanese	Kyoto4.0 (Kurohashi and Nagao, 1997)	42	97.4	98.7	99.3
Korean	Sejong (http://www.sejong.or.kr)	187	96.5	97.5	98.4
Portuguese	Floresta Sintá(c)tica/CoNLL06 (Afonso et al., 2002)	22	96.9	96.8	97.4
Russian	SynTagRus-RNC (Boguslavsky et al., 2002)	11	96.8	96.8	96.8
Slovene	SDT/CoNLL06 (Džeroski et al., 2006)	29	94.7	94.6	95.3
Spanish	Ancora-Cast3LB/CoNLL06 (Civit and Martí, 2004)	47	96.3	96.3	96.9
Swedish	Talbanken05/CoNLL06 (Nivre et al., 2006)	41	93.6	94.7	95.1
Turkish	METU-Sabanci/CoNLL07 (Oflazer et al., 2003)	31	87.5	89.1	90.2

Table 1: Data sets, number of language specific tags in the original treebank, and tagging accuracies for training/testing on the original (O) and the universal (U) tagset. Where applicable, we indicate whether the data set was extracted from the CoNLL 2006 (Buchholz and Marsi, 2006) or CoNLL 2007 (Nivre et al., 2007) versions of the corpora.

objective in publicly releasing this resource is to provide treebank and language specific experts a mechanism for refining these categories and the decisions we have made, as well as adding new treebanks and languages.

3. Experiments

To demonstrate the utility of the proposed universal POS tagset, we performed three sets of experiments. First, to provide a language comparison, we trained the same supervised POS tagging model on all of the above treebanks and evaluated the tagging accuracy on the universal POS tagset. Second, we used universal POS tags (automatically projected from English) as the starting point for unsupervised grammar induction, producing completely unsupervised parsers for several languages. Finally, we used the tagset in parser projection experiments where parallel data is used to transfer an English parser to new languages.

3.1. Language Comparisons

To compare POS tagging accuracies across different languages we trained a supervised tagger based on a trigram Markov model (Brants, 2000) on all treebanks. We chose this model for its fast speed and (close to) state-of-the-art accuracy without language specific tuning.[2]

Table 1 shows the results for all 25 treebanks when training/testing on the original (O) and universal (U) tagsets.

[2]Trained on the English PennTreebank this model achieves 96.7% accuracy when evaluated on the original 45 POS tags.

Overall, the variance on the universal tagset has been reduced by half (5.1 instead of 10.4). But of course there are still accuracy differences across the different languages. On the one hand, given a golden segmentation, tagging Japanese is almost deterministic, resulting in a final accuracy of above 99%. It is noteworthy that the accuracy on the two Japanese treebanks is almost the same when evaluating on the universal POS tags. For German, the two treebanks share the same fine-grained tagset, so the differences in accuracy are primarily due to domain effects and training set size variations. But again, when evaluating on the universal tagset, the results are almost identical. On the other hand, tagging Turkish, an agglutinative language with an average sentence length of 11.6 tokens, remains very challenging, resulting in an accuracy of only 90.2%.

Note that the best results are obtained by training on the original treebank categories and mapping the predictions to the universal POS tags at the end (O/U column). This is because the transition model based on the universal POS tagset is less informative. An interesting experiment would be to train the latent variable tagger of Huang et al. (2009) on the universal tagset. Their model automatically discovers refinements of the observed categories and could potentially find a tighter fit to the data than the one provided by the original, linguistically motivated tags.

3.2. Grammar Induction

We further demonstrate the utility of the universal POS tags in a grammar induction experiment. We combine the

Language	DMV	PGI	USR-G	USR-I	Transfer-G	Transfer-I
Danish	33.5	41.6	55.1	41.7	53.2	51.9
Dutch	37.1	45.1	44.0	38.8	67.6	66.9
German	35.7	-[3]	60.0	55.1	65.9	59.2
Greek	39.9	-[3]	60.3	53.4	73.9	72.5
Italian	41.1	-[3]	47.9	41.4	65.5	61.2
Portuguese	38.5	63.0	70.9	66.4	77.9	73.7
Spanish	28.0	58.4	68.3	43.3	58.0	51.4
Swedish	45.3	58.3	52.6	59.4	70.4	67.0

Table 2: Grammar induction results in terms of directed dependency accuracy. DMV and PGI use fine-grained gold POS tags, while USR-G and Transfer-G use gold universal POS tags and USR-I and Tranfer-I use automatically projected
universal POS tags.

cross-lingual part-of-speech projection framework of Das and Petrov (2011) with the grammar induction system of Naseem et al. (2010), to build parsers for languages without any labeled data resources. The tagger projection system assumes that the universal POS tag categories exist across languages and transfers the tags via word alignments. The grammar induction system uses a set of universal syntactic rules (USR), specified in terms of our universal POS tags, to constrain a probabilistic Bayesian model.

We present results on the same eight Indo-European languages as Das and Petrov (2011), so that we can make use of their automatically projected POS tags.[4] We used the treebanks released as part of the CoNLL-X shared task for all languages (Buchholz and Marsi, 2006). We only considered sentences of length 10 or less, after the removal of punctuations. Table 2 shows directed dependency accuracies for the DMV model of Klein and Manning (2004) and the PGI model of Berg-Kirkpatrick and Klein (2010) using fine-grained gold POS tags. For the USR model, we report results on gold universal POS tags (USR-G) and automatically induced universal POS tags (USR-I). The USR-I model falls short of the USR-G model, but has the advantage that it does not require any labeled data from the target language. Quite impressively, it does better than DMV for all languages, and is competitive with PGI, even though those models have access to fine-grained gold POS tags.

3.3. Parser Transfer

McDonald et al. (2011) present a parser projection system that relies heavily on our universal tagset. We replicate their baseline system here, which is very similar to the system of Zeman and Resnik (2008).

Statistical dependency parsers rely heavily on POS tag information. In fact, a delexicalized parser – a parser that has only non-lexical features – loses only 5-10% in accuracy compared to a state-of-the-art lexicalized parser. This observation combined with our universal part-of-speech tagset leads to the idea of direct transfer, i.e., directly parsing the target language with the source language parser. Because we use a mapping of the treebank specific part-of-speech tags to a common tagset, the performance of a such a system is easy to measure: simply parse the target language

[3]Not reported by Berg-Kirkpatrick and Klein (2010).

[4]The projected POS tags from their system are available at http://code.google.com/p/pos-projection/.

data set with a delexicalized parser trained on the source language data.

The last two columns of Table 2 show the performance of such a directly transferred parser using gold and projected universal POS tags. Perhaps somewhat surprisingly, this simplistic approach actually outperforms state-of-the-art unsupervised grammar induction systems, and highlights the utility and information contained in our coarse universal POS tags.

4. Conclusions

We proposed a POS tagset consisting of twelve categories that exists across languages and developed a mapping from 25 language specific tagsets to this universal set. We demonstrated experimentally that the universal POS categories generalize well across language boundaries on an unsupervised grammar induction task, as well as a parser transfer task, giving competitive parsing accuracies without relying on gold POS tags. The tagset and mappings are available for download at **http://code.google.com/p/universal-pos-tags/**

Acknowledgements

We would like to thank Joakim Nivre for allowing us to use a preliminary tagset mapping developed in McDonald and Nivre (2007). The second author was supported in part by NSF grant IIS-0844507.

5. References

A. Abeillé, L. Clément, and F. Toussenel. 2003. Building a Treebank for French. In Abeillé (Abeillé, 2003), chapter 10.

A. Abeillé, editor. 2003. *Treebanks: Building and Using Parsed Corpora*. Kluwer.

I. Aduriz, M. J. Aranzabe, J. M. Arriola, A. Atutxa, A. Diaz de Ilarraza, A. Garmendia, and M. Oronoz. 2003. Construction of a Basque dependency treebank. In *Proc. of the Workshop on Treebanks and Linguistic Theories*.

S. Afonso, E. Bick, R. Haber, and D. Santos. 2002. Floresta sintá(c)tica: a treebank for Portuguese. In *Proc. of LREC*.

T. Berg-Kirkpatrick and D. Klein. 2010. Phylogenetic grammar induction. In *Proc. of ACL*.

I.M. Boguslavsky, L.L. Iomdin, I.S. Chardin, and L.G. Kreidlin. 2002. Development of a dependency treebank for russian and its possible applications in nlp. In *Proc. of LREC*.

A. Böhmová, J. Hajič, E. Hajičová, and B. Hladká. 2003. The PDT: a 3-level annotation scenario. In Abeillé (Abeillé, 2003), chapter 7, pages 103–127.

S. Brants, S. Dipper, S. Hansen, W. Lezius, and G. Smith. 2002. The TIGER Treebank. In *Proc. of the Workshop on Treebanks and Linguistic Theories*.

T. Brants. 1995. Tagset reduction without information loss. *Proc. of ACL*.

T. Brants. 2000. TnT - a statistical part-of-speech tagger. In *Proc. of ANLP*.

S. Buchholz and E. Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proc. of CoNLL*.

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

K. Chen, C. Luo, M. Chang, F. Chen, C. Chen, C. Huang, and Z. Gao. 2003. Sinica treebank: Design criteria, representational issues and implementation. In Abeillé (Abeillé, 2003), chapter 13, pages 231–248.

C. Christodoulopoulos, S. Goldwater, and M. Steedman. 2010. Two decades of unsupervised POS induction: How far have we come? In *Proc. of EMNLP*.

M. Civit and M.A. Martí. 2004. Building cast3lb: A spanish treebank. *Research on Language & Computation*, 2(4):549–574.

S. B. Cohen, D. Das, and N. A. Smith. 2011. Unsupervised structure prediction with non-parallel multilingual guidance. In *Proc. of EMNLP*.

D. Csendes, J. Csirik, T. Gyimóthy, and A. Kocsor. 2005. *The Szeged Treebank*. Springer.

D. Das and S. Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proc. of ACL-HLT*.

J. DeNero and J. Uszkoreit. 2011. Inducing sentence structure from parallel corpora for reordering. In *Proc. of EMNLP*.

M. Dickinson and C. Jochim. 2008. A simple method for tagset comparison. *Proc. of LREC*.

P. Dienes and C. Oravecz. 2000. Bottom-up tagset design from maximally reduced tagset. In *Proc of the COLING Workshop on Linguistically Interpreted Corpora*.

M. Ariel Dominguez and G. Infante-Lopez. 2008. Searching for part of speech tags that improve parsing models. *Lecture Notes in Computer Science*.

S. Džeroski, T. Erjavec, N. Ledinek, P. Pajas, Z. Žabokrtsky, and A. Žele. 2006. Towards a Slovene dependency treebank. In *Proc. of LREC*.

T. Erjavec. 2004. MULTEXT-East version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proc. of LREC*.

N. Evans and S. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(05).

K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proc. of ACL*.

J. Hajič, O. Smrž, P. Zemánek, J. Šnaidauf, and E. Beška. 2004. Prague Arabic dependency treebank: Development in data and tools. In *Proc. of NEMLAR*.

Z. Huang, V. Eidelman, and M. Harper. 2009. Improving simple bigram HMM part-of-speech tagger by latent annotation. In *Proc. of NAACL-HLT*.

Y. Kawata and J. Bartels. 2000. Stylebook for the Japanese treebank in VERBMOBIL.

M.J. Kim. 2002. Does Korean have adjectives? *MIT Working Papers in Linguistics*, 43:71–89.

D. Klein and C. D. Manning. 2004. Corpus-based induction of syntactic structure: models of dependency and constituency. In *Proc. of ACL*.

M.T. Kromann, L. Mikkelsen, and S.K. Lynge. 2003. Danish Dependency Treebank. In *Proc. of the Workshop on Treebanks and Linguistic Theories*.

S. Kurohashi and M. Nagao. 1997. Kyoto University text corpus project. In *Proc. of ANLP*.

C. Manning. 2011. Part-of-speech tagging from 97% to

100%: is it time for some linguistics? *Proc. of CICLing*.

M. P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19.

M. A. Martí, M. Taulé, L. Màrquez, and M. Bertran. 2007. CESS-ECE: A multilingual and multilevel annotated corpus. Available for download from: http://www.lsi.upc.edu/∼mbertran/cess-ece/.

R. McDonald and J. Nivre. 2007. Characterizing the errors of data-driven dependency parsing models. In *Proc. of EMNLP-CoNLL*.

R. McDonald, S. Petrov, and K. Hall. 2011. Multisource transfer of delexicalized dependency parsers. In *Proc. of EMNLP*.

S. Montemagni, F. Barsotti, M. Battista, N. Calzolari, O. Corazzari, A. Lenci, A. Zampolli, F. Fanciulli, M. Massetani, R. Raffaelli, R. Basili, M. T. Pazienza, D. Saracino, F. Zanzotto, N. Nana, F. Pianesi, and R. Delmonte. 2003. Building the Italian Syntactic-Semantic Treebank. In Abeillé (Abeillé, 2003), chapter 11, pages 189–210.

T. Naseem, B. Snyder, J. Eisenstein, and R. Barzilay. 2009. Multilingual part-of-speech tagging: Two unsupervised approaches. *JAIR*, 36.

T. Naseem, H. Chen, R. Barzilay, and M. Johnson. 2010. Using universal linguistic knowledge to guide grammar induction. In *Proc. of EMNLP*.

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

J. Nivre, J. Nilsson, and J. Hall. 2006. Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation. In *Proc. of LREC*.

J. Nivre, J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel, and D. Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proc. EMNLP-CoNLL*.

K. Oflazer, B. Say, D. Zeynep Hakkani-Tür, and G. Tür. 2003. Building a Turkish treebank. In Abeillé (Abeillé, 2003), chapter 15, pages 261–277.

M. Palmer, N. Xue, F. Xia, F. Chiou, Z. Jiang, and

M. Chang. 2007. Chinese Treebank 6.0. Technical report, Linguistic Data Consortium, Philadelphia. P.

Prokopidis, E. Desypri, M. Koutsombogera, H. Papageorgiou, and S. Piperidis. 2005. Theoretical and practical issues in the construction of a Greek dependency treebank. In *Proc. of the Workshop on Treebanks and Linguistic Theories*.

O. Rambow, B. Dorr, D. Farwell, R. Green, N. Habash, S. Helmreich, E. Hovy, L. Levin, K. J. Miller, T. Mitamura, Reeder F., and A. Siddharthan. 2006. Parallel syntactic annotation of multiple languages. In *Proc. of LREC*.

L. Shen, G. Satta, and A. Joshi. 2007. Guided learning for bidirectional sequence classification. In *Proc. of ACL*.

K. Simov, P. Osenova, S. Kolkovska, E. Balabanova, D. Doikoff, K. Ivanova, A. Simov, and M Kouylekov. 2002. Building a Linguistically Interpreted Corpus of Bulgarian: the BulTreeBank. In *Proc. of LREC*.

W. Skut, B. Krenn, T. Brants, and H. Uszkoreit. 1997. An annotation scheme for free word order languages. In *Proc. of ANLP*.

N. A. Smith and J. Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proc. of ACL*.

B. Snyder, T. Naseem, J. Eisenstein, and R. Barzilay. 2009. Adding more languages improves unsupervised multilingual part-of-speech tagging: A Bayesian non-parametric approach. In *Proc. of NAACL*.

K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of HLT-NAACL*.

L. Van der Beek, G. Bouma, R. Malouf, and G. Van Noord. 2002. The Alpino dependency treebank. *Language and Computers*, 45(1):8–22.

C. Xi and R. Hwa. 2005. A backoff model for bootstrapping resources for non-English languages. In *Proc. of HLT-EMNLP*.

D. Yarowsky and G. Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proc. of NAACL*.

D. Zeman and P. Resnik. 2008. Cross-language parser adaptation between related languages. In *NLP for Less Privileged Languages*.

D. Zeman. 2008. Reusable tagset conversion using tagset drivers. *Proc. of LREC*.

ADJ	ADP	ADV	CONJ	DET	NOUN	NUM	PRON	PRT	VERB	X	.
A-	P-	D-	C-		N-, Z-	Q-	S-, SD, SR	F-, Fl, FN	VC, VI, VP, _	–, I-, Y-,	G-
Omitted for space reasons. 64 tags. See http://code.google.com/p/universal-pos-tags/.											
A, Af, Am, An, H, Hf, Hm, Hn, Md,	R	D, Dd, Dl, Dm, Dq, Dt	Cc, Cp, Cr, Cs		My, N, Nc, Nm, Np	Mc	P, Pc, Pd, Pf, Pi, Pn, Pp, Pr, Ps	Ta, Te, Tg, Ti, Tm, Tn, Tv, Tx	V, Vii, Vni, Vnp, Vpi, Vpp, Vxi, Vyp	I	Punct
ao, aq	sp	rg, rn	cc, cs	da, dd, de, di, dn, dp, dr, dt	nc, np	W, Z, Zm, Zp, w, z, zm, zp	p0, pd, pi, pn, pp, pr, pt, px		va, vm, vs	I, i, _	Fa, Fc, Fd, Fe Fg, Fh, Fi, Fp, Fs, Fx, Fz, fa fc, fg, fp
JJ	P	AD	CC, CS	DT	NN, NR, NT	CD, M, OD	PN	AS, DEC, DEG, DER, DEV, ETC, LC, MSP, SP	VA, VC, VE, VV	BA, FW, IJ, LB, ON, SB, X	PU
Omitted for space reasons. 294 tags. See http://code.google.com/p/universal-pos-tags/.											
2, A, C, G, M, U	F, S, V	b, g	,, ^		N	*, =, ?, a, d, h, k, l, n, o, r, u, v, w, y, }	1, 4, 5, 6, 7, 8, 9, D, E, H, J, K, L, O, P, R, W, Z	T	B, c, e, f, i, m, p, s	@, I, X, x	:
AC, AN, AO	SP	RG	CC, CS		NC, NP		PC, PD, PI, PO, PP, PT		VA, VE	I, U, XA, XF, XP, XR, XS, XX	
Adj	Prep	Adv	Conj	Art	N	Num	Pron		V	Int, Misc	Punc
JJ, JJR, JJS	IN	RB, RBR, RBS, WRB	CC	DT, EX, PDT, WDT	NN, NNP, NNPS, NNS	CD	PRP, PRP$, WP, WP$	POS, RP, TO	MD, VB, VBD, VBG, VBN, VBP, VBZ	FW, LS, SYM, UH	#, $, ”, ,, -LRB- ., :, “, -RRB-
ADJ, ADJWH	P, P+D, P+PRO, P+PRON	ADV, ADVWH	CC, CS	DET, DETWH	NC, NPP		CLO, CLR, CLS, PRO, PROREL, PROWH	PREF	V, VIMP, VINF, VPP, VPR, VS	ET, I	PONCT
ADJA, ADJD	APPO, APPR, APPRART, APZR	ADV	KOKOM, KON, KOUI, KOUS	ART	NE, NN, NNE	CARD	PDAT, PDS, PIAT, PIS, PPER, PPOSAT, PPOSS, PRELAT, PRELS,	PTKA, PTKANT, PTKNEG, PTKVZ, PTKZU	VAFIN, VAIMP, VAINF, VAPP, VMFIN, VMINF, VMPP, VVFIN, VVIMP, VVINF, VVIZU, VVPP	FM, ITJ, TRUNC, XY	$(, $,, $.
ADJA, ADJD	APPO, APPR, APPRART, APZR	ADV	KOKOM, KON, KOUI, KOUS	ART	NE, NN, NNE	CARD	PDAT, PDS, PIAT, PIS, PPER, PPOSAT, PPOSS, PRELAT, PRELS,	PTKA, PTKANT, PTKNEG, PTKVZ, PTKZU	VAFIN, VAIMP, VAINF, VAPP, VMFIN, VMINF, VMPP, VVFIN, VVIMP, VVINF, VVIZU, VVPP	FM, ITJ, TRUNC, XY	$(, $,, $.

	CONJ	DET	NOUN	NUM	PRON	PRT	VERB	X	.
	CjCo, CjSb	AtDf, AtId	NoCm, NoPr	DATE, DIG, ENUM, NmCd, NmCt, NmMl, NmOd	PnDm, PnId, PnIr, PnPe, PnPo, PnRe, PnRi	PtFu, PtNg, PtOt, PtSj	VbIs, VbMn	COMP, INIT, LSPLIT, RgAbXx, RgAnXx, RgFwOr, RgFwTr	PUNCT
i, p,	Cc, Cs	Tf, Ti	Nc, Np	Mc, Md, Mf, Mo	Pd, Pg, Pi, Pp, Pq, Pr, Ps, Px, Py		Va, Vm	I, Io, Oh, Oi, On, X, Y, Z	SPUNCT, WPUNCT
	C	DD, DE, DI, DR, DT, RD, RI	S, SP, SW	N, NO	PD, PI, PP, PQ, PR, PT		V	I, SA, X	PU
	CON	DA, DN, DP	LOC, NA, NC, NF, NP, NT, NV, ORG, PER, PN, SN SNN, SNP, SNS	NUM	NR	COP, PC PCO, PF, PS	AUX, PV, SV, V	INT, X	(,), " " SYM
Omitted for space reasons. 80 tags. See http://code.google.com/p/universal-pos-tags/.									
Omitted for space reasons. 187 tags. See http://code.google.com/p/universal-pos-tags/.									
	conj-c, conj-s	art	n, pp, prop	num	pron-det, pron-indp, pron-pers		v-fin, v-ger, v-inf, v-pcp, vp	ec, in	?, punc
	C		N	M	P	Q	V	I	X
Omitted for space reasons. 29 tags. See http://code.google.com/p/universal-pos-tags/.									
	cc, cs	da, dd, de, di, dn, dp, dt	nc, np	Zm, Zp, w, z	p0, pd, pe, pi, pn, pp, pr, pt, px		va, vm, vs	X, Y, i	Fa, Fc, Fd, Fe, Fg, Fh, Fi, Fp, Fs, Fx, Fz
	UK, ++		AN, MN, NN, PN, VN	EN, RO	PO	IM	AV, BV, FV, GV HV, KV, MV, QV , SP, SV, TP, VV, WV	ID, XX, YY	I?, IC, IG, IK, IP, IQ, IR, IS, IT, IU, PU
	Conj	Det	NFutPart, NInf, NPastPart, NPresPart, Noun, Prop	Card, Distrib, Num, Ord, Range, Real	DemonsP, PersP, Pron, QuesP, ReflexP	Dup	Verb, Zero	Interj	Punc, Ques

ppings from language-specific part-of-speech tags to our universal part-of-speech tags.

Language
Arabic PADT/CoNLL07 (Hajič et al., 2004)
Basque Basque3LB/CoNLL07 (Aduriz et al., 2003)
Bulgarian BTB/CoNLL06 (Simov et al., 2002)
Catalan CESS-ECE/CoNLL07 (Martí et al., 2007)
Chinese Penn ChineseTreebank 6.0 (Palmer et al., 2007)
Chinese Sinica/CoNLL07 (Chen et al., 2003)
Czech PDT/CoNLL07 (Böhmová et al., 2003) (Böhmová et al., 2003)
Danish DDT/CoNLL06 (Kromann et al., 2003)
Dutch Alpino/CoNLL06 (Van der Beek et al., 2002)
English PennTreebank (Marcus et al., 1993)
French FrenchTreebank (Abeillé et al., 2003)
German Tiger/CoNLL06 (Brants et al., 2002)
German Negra (Skut et al., 1997)

Language	ADJ	ADP	ADV
Greek GDT/CoNLL07 (Prokopidis et al., 2005)	Aj	AsPpPa, AsPpSp	Ad
Hungarian Szeged/CoNLL07 (Csendes et al., 2005)	Af	St	Rd, Rg, R Rl, Rm, R Rq, Rr, Rv, Rx
Italian ISST/CoNLL07 (Montemagni et al., 2003)	A, AP	E	B
Japanese Kyoto4.0 (Kurohashi and Nagao, 1997)	ADJ, ADN, PA, SAN, SAP		ADV
Japanese Verbmobil/CoNLL06 (Kawata and Bartels, 2000)			
Korean Sejong (http://www.sejong.or.kr)			
Portuguese Floresta Sintá(c)tica/CoNLL06 (Afonso et al., 2002)	adj	prp	adv
Russian SynTagRus-RNC (Boguslavsky et al., 2002)	A	S	R
Slovene SDT/CoNLL06 (Džeroski et al., 2006)			
Spanish Ancora-Cast3LB/CoNLL06 (Civit and Martí, 2004)	ao, aq	sn, sp	rg, rn
Swedish Talbanken05/CoNLL06 (Nivre et al., 2006)	AJ	PR	AB
Turkish METU-Sabanci/CoNLL07 (Oflazer et al., 2003) (Oflazer et al., 2003)	AFutPart, APastPart, APresPart, Adj	Postp	Adv

Table 3: The proposed ma