Cross-lingual tagger evaluation without test data

Željko Agić
IT University of Copenhagen
zeag@itu.dk

Barbara Plank
University of Groningen
b.plank@rug.nl

Anders Søgaard
University of Copenhagen
soegaard@di.ku.dk

Abstract

We address the challenge of cross-lingual POS tagger evaluation in absence of manually annotated test data. We put forth and evaluate two dictionary-based metrics. On the tasks of accuracy prediction and system ranking, we reveal that these metrics are reliable enough to approximate test set-based evaluation, and at the same time lean enough to support assessment for truly low-resource languages.

1 Introduction

Cross-lingual learning of NLP models is currently in an evaluation impasse. While we can create reliable cross-lingual taggers and parsers for hundreds of low-resource languages (Agić et al., 2016), we can only evaluate our models for languages where some hand-annotated test data is available. The requirement for the uniformity of annotations (McDonald et al., 2013) further strengthens the constraint. The set of languages with readily available test data is very exclusive. Namely, they are the resource-rich languages from the Universal Dependencies project (Nivre et al., 2015).[1]

Recent works have suggested to evaluate cross-lingual approaches *by proxy*, e.g., by using crowdsourced tag dictionaries (Li et al., 2012; Agić et al., 2015). In these works, though, the validity of assessment by using tag dictionaries is left completely unaddressed.

Contributions. Our work poses the question: How adequate are tag dictionaries for evaluating POS taggers for low-resource languages? Across 25 languages, we compare the POS tagger rankings induced by evaluation against dictionaries to those induced by evaluation on manually annotated gold standards. We select the best out of five competitive taggers for 14 out of 25 languages. We also consider to what extent we can predict true tagging scores. We find that as little as the 100 most frequent tokens with corresponding POS tags suffice to provide reliable estimates of true scores. Finally, we introduce a novel metric that presumes nothing but an English tag dictionary and a small bilingual dictionary for the target language. We also find this metric to be a relatively robust estimator for tagging accuracy. It finds the best tagger for 11 out of 20 languages.

Our code and data are freely available.[2]

2 Metrics

In cross-lingual learning work, it is common to evaluate POS taggers for accuracy by using test data annotated by human experts. For a test set T of n word-tag pairs (w_i, t_i) and its tagging $\hat{T}$, we define the true accuracy A_{true} as:

$$A_{\text{true}}(\hat{T}, T) = \frac{|\{(w_i, \hat{t}_i) \in \hat{T} \mid \hat{t}_i = t_i\}_{i=1}^{n}|}{|\hat{T}|}$$

$$T = \{(w_i, t_i)\}, \hat{T} = \{(w_i, \hat{t}_i)\}, 1 \leq i \leq n$$

Obviously this metric can only be computed when test data is available, which is not the case for the vast majority of the world's languages. Note that while we use the term *true* accuracy, the adequacy of the metric depends on how representative the annotated data is of the underlying distribution.

Drawing from Li et al. (2012)—who compared Wiktionaries to gold dictionaries extracted from the tagger training sets—Agić et al. (2015) propose an approximate metric in absence of test data T. They apply it to 10 low-resource languages by

[1] http://universaldependencies.org/

[2] Wiktionaries included, https://bitbucket.org/lowlands/release.

using Wiktionaries ranging from only 50 to more than 20k dictionary entries. We take their metric as our starting point.

Soft accuracy. Given a dictionary $\mathcal{D}$ whose entries are word forms with their ambiguous taggings $(w, D_w = \{t_1^w, ..., t_k^w\})$, we express the approximate or soft accuracy A_{soft} as:

$$A_{\text{soft}}(\hat{T}, \mathcal{D}) = \frac{|\{(w_i,\hat{t}_i)\in\hat{T}|(w_i,D_{w_i})\in\mathcal{D}\wedge\hat{t}_i\in D_{w_i}\}_{i=1}^{n}|}{|\{(w_i,\hat{t}_i)\in\hat{T}|(w_i,D_{w_i})\in\mathcal{D}\}_{i=1}^{n}|}$$

In absence of true tags t_i, we ambiguously tag T using the tags from $\mathcal{D}$, but only for the tokens w_i that are covered by the dictionary: $(w_i, D_{w_i}) \in \mathcal{D}$. We then count the tagger output $\hat{t}_i$ as correct iff it is warranted by the dictionary: $\hat{t}_i \in D_{w_i}$.

Problems. Crowd-sourced dictionaries can suffer from limited coverage and poor quality. We counter the first issue by covering the most frequent words. We distinguish between A_{soft} with frequency information (+freq), using the m most frequent words, or without frequency information ($-$freq), using m random words.

Tag lists D_i can also be deficient: They can be missing certain tags, or contain incorrect tags, or both. For example, the Croatian Wiktionary only notes the NOUN tagging of *igra* (en. *game*), but in reality the word form also has a VERB tagging (en. *to play*, third person singular).

We can gauge the quality of $\mathcal{D}$ in presence of a high-quality dictionary $\mathcal{G} = \{(w_i, G_i)\}_{i=1}^{|\mathcal{G}|}$ which we can induce from a training set:

$$\text{precision}(\mathcal{D}, \mathcal{G}) = \sum_{i=1}^{|\mathcal{D}|} \frac{|\{D_i \cap G_i\}|}{|\{t \in D_i\}|}$$

$$\text{recall}(\mathcal{D}, \mathcal{G}) = \sum_{i=1}^{|\mathcal{D}|} \frac{|\{D_i \cap G_i\}|}{|\{t \in G_i\}|}$$

Namely, for each word w_i covered by both $\mathcal{D}$ and $\mathcal{G}$, we check how many tags D_i and G_i intersect, and then use the intersection to estimate dictionary precision and recall.

Translated dictionaries. With low-resource languages, we cannot presume the availability of tag dictionaries. However, we often have high-quality bilingual dictionaries with translations of common words into a resource-rich language such as English. With these in place, we can "translate" the English dictionary into a low-resource language and exploit the resulting $\mathcal{D}_{\text{trans}}$ in the evaluation for A_{soft}. We implement a very simple form of dictionary lookup-based translation, whereby all words in the English word-tag dictionary are replaced by target-language words through bilingual dictionaries.

We expect this bilingual dictionary-based soft metric A_{trans} to suffer from the same coverage and quality problems as A_{soft}, and to introduce additional "translation noise" on top of that. We maintain that both metrics can still be reliable estimators of tagging accuracy for truly low-resource languages in absence of annotated test data.

3 Experiments

We perform two sets of experiments:

i) **numerical score prediction**, where we evaluate the approximate metrics A_{soft} and A_{trans} as estimators of the true POS tagging accuracies A_{true}, and

ii) **rank prediction**, where we test how well do A_{soft} and A_{trans} perform in ranking several POS taggers relative to A_{true}.

In numerical score prediction, we evaluate the taggers using all three metrics, and establish empirical relations between dictionary quality and size, and the observed scores.

In rank prediction, we rank five POS taggers using A_{true}, and then attempt to replicate the ranking using A_{soft} and A_{trans}. We express the quality of predicted rankings using precision (P@1) and Kendall's τ_b statistic (Knight, 1966).

Data. We train and test our taggers on data from UD version 1.2 (Nivre et al., 2015). We intersect this collection with the dictionaries we make available for this experiment: 9 of the Wiktionaries come from Li et al. (2012), and we collect 16 new on top of that. Thus, we experiment with a total of 25 languages from the UD. We refer to the 9 languages of Li et al. (2012) as development languages. To make the Wiktionaries and the UD data compatible, we map all POS tags to the tagset by Petrov et al. (2012).

We estimate the frequencies for the +freq variants of the soft metrics by using the multilingual Bible corpus by Christodouloupoulos and Steedman (2014) and the Watchtower corpus (Agić et al., 2016) combined.

We translate the English Wiktionary from Li et al. (2012) by using bilingual dictionaries from Wiktionary to obtain $\mathcal{D}_{\text{trans}}$ for 20 languages.[3]

[3]We choose the English Wiktionary rather than the En-

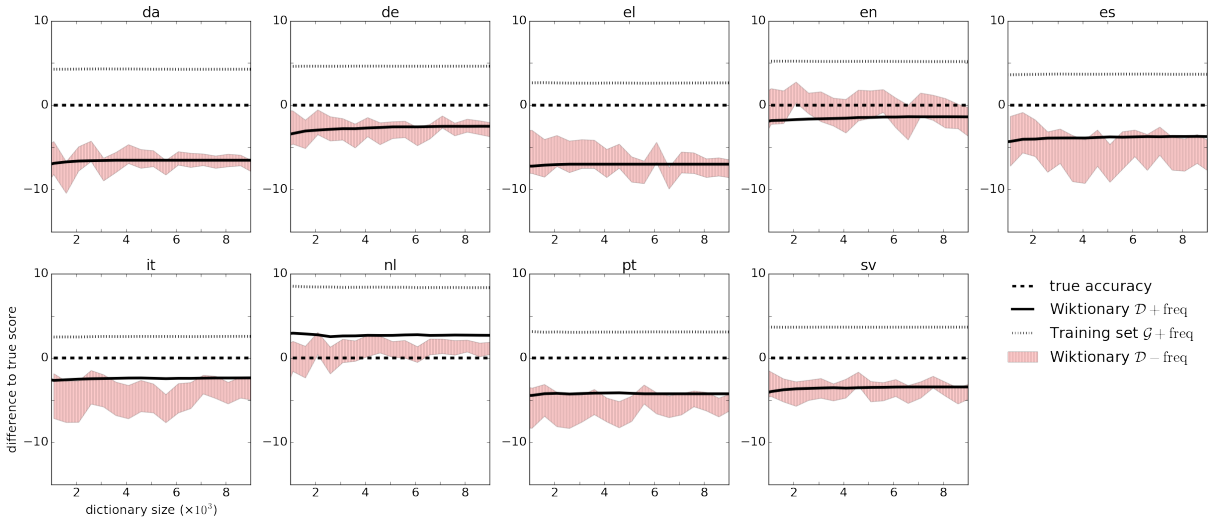

Figure 1: Impact of dictionary size and frequency usage (−freq, +freq) on numerical score prediction for nine development languages using the TnT tagger. The shaded regions represent 95% confidence intervals for $\mathcal{D}-\text{freq}$. The −freq dictionaries are randomly sampled 100 times for each size step, and the steps range 100–10,000 entries both for −freq and +freq.

Taggers. We experiment with five POS taggers, all run with their default settings:

a) bi-LSTM tagger (Plank et al., 2016),
b) CRF++ (Kudo, 2005),
c) MarMoT (Mueller et al., 2013),
d) TnT (Brants, 2000), and
e) TreeTagger (Schmid, 1994).

3.1 Results

Score prediction. Here, we discuss how well our metric A_{soft} performs in guessing the true tagger accuracies by using the Wiktionaries.

Figure 1 reveals that even large Wiktionaries do not make for good accuracy estimators if they do not exploit the frequencies. We see the evidence for that in the very wide confidence intervals in our Wiktionary sampling. In contrast, even the smallest of frequency-aware Wiktionaries prove to be much more reliable. They can contain as little as 100 entries, especially if their tagging quality is high. For example, a bad sample of 6k Spanish (es) words and tags might underestimate A_{true} by 10 points, while using the 100 most frequent Spanish words get us as close as -4 points even with erroneous tags.

We observe high negative correlations of Wiktionary F_1 scores (Pearson's $\rho = -0.58$) and test set coverages ($\rho = -0.60$) with the quality of accuracy estimation, expressed as absolute difference of the two scores $|A_{\text{true}} - A_{\text{soft}}|$ for the data in Figure 1. In simpler terms: The higher i) the intrinsic quality of the Wiktionary and ii) its coverage, the better the score estimation. There, the Wiktionaries are intrinsically evaluated with respect to the training set dictionaries. We also note that the noisy Wiktionaries ($\mathcal{D}$) tend to underestimate A_{true}, while the more reliable gold dictionaries ($\mathcal{G}$) overestimate.

The translation-based metric A_{trans} approximates the true scores better than A_{soft} for 7/20 languages, and is more stable across languages as all $\mathcal{D}_{\text{trans}}$ originate from English (en). See Table 1 for the results on all 25 languages.

Rank prediction. In system ranking, we try to select the best tagger for a given language through our metrics. We note the task is rather hard as all the taggers score very close to one another. Still, we manage to find the best tagger for 14/25 languages with A_{soft}, and for 11/20 with A_{trans}.

For some languages, even in spite of Wiktionary deficiency, we manage to i) select the best tagger and to ii) improve the true score prediction through translation from English. For example, the high quality of Bulgarian (bg) Wiktionary is outweighed by the high coverage of its $\mathcal{D}_{\text{trans}}$, and there A_{trans} significantly improves the prediction. For Farsi (fa), we improve both the score predic-

glish UD training set due to much higher coverage in spite of lower precision: $F_1 = 18.51$ for the Wiktionary translations ($\mathcal{D}_{\text{trans}}$), compared to $F_1 = 13.22$ for the UD training set translations ($\mathcal{G}_{\text{trans}}$) over 20 languages.

	Wiktionary quality						*Metrics evaluation*						
	$\mathcal{D}$			$\mathcal{D}_{\text{trans}}$			A_{true}	A_{soft}			A_{trans}		
	$\lvert\mathcal{D}\rvert$	P	R	$\lvert\mathcal{D}\rvert$	P	R	$\bar{A}_{\text{true}}$	$\bar{A}_{\text{soft}}$	P@1	τ_b	$\bar{A}_{\text{trans}}$	P@1	τ_b
Bulgarian (bg)	3	93.58	3.54	15	59.33	7.65	97.45±1.14	89.73±0.20	0	-0.2	**95.54**±0.18	0	-0.2
Czech (cs)	14	98.77	4.82	23	62.35	5.59	97.88±1.00	**94.74**±0.82	1	0.6	93.47±0.19	0	0.2
* Danish (da)	23	83.89	19.00	15	55.42	12.21	96.07±1.03	**88.94**±0.51	1	0.6	87.54±0.51	0	0.4
* German (de)	63	94.97	23.19	46	63.20	14.79	95.02±0.42	**92.48**±0.21	1	0.4	76.77±0.24	1	0.2
* Greek (el)	22	87.99	18.72	21	56.50	10.85	96.97±1.22	**89.45**±0.47	1	1.0	78.28±0.32	1	0.4
* English (en)	388	69.88	65.97	–	–	–	95.39±1.07	93.15±0.26	0	-0.4	–	–	–
* Spanish (es)	240	85.00	40.20	31	67.27	17.00	96.22±0.38	**91.91**±0.65	0	-0.2	79.39±0.36	1	0.4
Basque (eu)	1	90.43	1.49	–	–	–	95.08±1.33	74.90±1.25	1	0.8	–	–	–
Farsi (fa)	4	87.87	11.89	1	56.22	1.43	96.35±0.73	90.46±0.45	0	-0.2	**94.05**±0.54	1	0.6
Finnish (fi)	104	88.41	8.52	45	53.70	6.38	94.72±2.24	79.72±1.26	1	1.0	**90.51**±0.44	1	0.8
French (fr)	17	88.70	7.49	36	67.55	18.55	96.47±0.62	48.08±1.03	0	0.2	**79.30**±0.23	0	0.2
Irish (ga)	6	85.73	12.54	–	–	–	92.77±0.87	91.97±2.88	0	0.2	–	–	–
Ancient Greek (grc)	5	94.13	2.46	–	–	–	91.97±2.88	74.62±0.96	1	1.0	–	–	–
Hebrew (he)	4	83.12	5.04	7	58.37	5.06	95.69±1.15	**86.23**±0.65	1	0.6	79.84±0.57	1	0.4
Hindi (hi)	2	89.79	4.19	2	61.03	3.81	97.97±0.73	**81.25**±1.03	0	0.2	80.16±0.50	0	-0.4
Croatian (hr)	21	92.03	12.76	6	55.44	2.43	95.32±1.06	89.81±0.46	0	0.4	**94.41**±0.76	0	-0.2
Hungarian (hu)	14	84.01	15.29	17	49.78	10.75	92.46±2.35	**86.22**±2.43	0	-0.2	73.04±1.14	1	0.4
* Italian (it)	494	79.03	65.29	29	63.32	19.62	97.53±0.61	**94.58**±0.46	1	0.6	85.51±0.23	0	0.2
Latin (la)	30	68.15	7.80	–	–	–	91.24±2.27	68.24±2.11	1	1.0	–	–	–
* Dutch (nl)	55	83.67	35.25	29	57.45	16.92	92.38±2.11	**92.85**±0.74	0	0.2	86.58±0.76	1	0.6
Norwegian (no)	47	89.51	6.94	11	55.32	7.48	97.67±0.55	33.99±0.11	0	0.2	**87.33**±0.14	0	0.2
Polish (pl)	6	92.97	3.70	22	53.91	8.50	95.55±1.35	**87.97**±1.17	1	0.8	81.86±0.28	1	0.6
* Portuguese (pt)	42	90.55	18.38	26	62.10	17.98	97.22±0.69	**92.39**±0.22	1	0.6	82.46±0.27	0	0.2
Romanian (ro)	7	82.29	16.95	15	49.65	16.64	89.59±2.26	83.59±1.50	1	1.0	**84.24**±0.74	1	1.0
* Swedish (sv)	91	85.84	48.32	29	53.89	16.60	96.22±0.92	**92.14**±0.77	1	0.4	83.32±0.49	1	0.4
Mean	–	86.81	18.38	–	58.09	11.01	95.25±0.89	83.27±5.69	14/25	0.42	86.40±2.89	11/20	0.27

Table 1: Wiktionary size and quality, and metrics evaluation. The dictionary sizes $|\mathcal{D}|$ are $\times 10^3$ entries. Wiktionaries are evaluated for precision (P) and recall (R) against the respective UD training set dictionaries ($\mathcal{G}$). In metrics evaluation, scores are obtained by using the full Wiktionaries, averaged ($\bar{A}$) over five POS taggers. *: development languages, with Wiktionaries by Li et al. (2012). ±: 95% confidence intervals; **bold**: best score estimates, i.e., lowest differences to true scores $|A_{\text{true}} - A_{\text{soft}}|$.

tion and the tagger selection.

Through Kendall's τ_b statistic, we rate the quality of the entire rankings, not just of guessing the best out of five taggers. We find that the true and the estimated rankings are statistically dependent at $p < 0.05$ for all languages. We also find that the taggers are easier to rank when the true scores are lower and further apart. For example, the French (fr) and Spanish (es) taggers are hard to rank as they all score very close to one another, while we easily rank the taggers for Greek (el), Basque (eu), Polish (pl), or Romanian (ro). We argue that such ranking behavior favors evaluation for low-resource languages, where insufficient data is very likely to cause even greater disparity between different POS taggers.

3.2 Discussion

Sources of POS tags. Our work aims at supporting cross-lingual POS tagger evaluation. Why did we then evaluate the metrics on outputs of fully supervised taggers? In short, because higher tagging scores are *harder* to estimate.

We experimented with: i) fully supervised taggers, ii) actual cross-lingual taggers from Agić et al. (2016), for which $\bar{A}_{\text{true}} = 70.56$, and iii) artificial corruption of gold POS tags.

In artificial data corruption for the development languages, we found that the score prediction error correlates with the true score ($\rho = 0.54$). For the corruption, we created 20 samples of $A_{\text{true}} \in [0, 1]$ for each language with a 0.05 increment. Further, we evaluated A_{soft} on the cross-lingual taggers. There, we singled out the best taggers for 13/21 intersecting languages, or for 2 languages more than over fully supervised taggers (11/21). With translated dictionaries, i.e., through A_{trans}, we scored 13/20 (also +2 languages).

For these reasons, we decided to show how our metrics perform in the most difficult case. Here,

the additional experiments with different sources of POS tags show that the metrics easily scale down to evaluating cross-lingual taggers for low-resource languages.

Held-out data. Annotating a handful of test sentences could serve as an alternative to dictionary-based evaluation. We find that ~55±27 sentences are needed on average to reach the system ranking accuracy of A_{trans} for our 20 languages. However, the option of annotating test data might not be feasible for many low-resource languages, while Wiktionaries are currently readily available for more than 300 languages. We also note that the required sample size is negatively correlated with tagging accuracy ($\rho = -0.63$): the lower the tagger accuracy, the more sentences we need to reasonably estimate it.

4 Related work

Li et al. (2012) gauge 9 Wiktionaries against gold dictionaries to strengthen the argument for their weakly-supervised tagger. Agić et al. (2015) use 10 Wiktionaries to extend a cross-lingual tagger evaluation to languages without test sets, but they do so indiscriminately. Their Wiktionaries range from only 50 to more than 20k random entries. To the best of our knowledge, research on evaluating POS taggers in absence of manually annotated test data is novel to our work.

We collected 16 new Wiktionaries on top of the 9 provided by Li et al. (2012) for our experiment. Recently, larger Wiktionary datasets[4] have been made available, enabling further experiments with cross-lingual tagging. The dataset of Sylak-Glassman et al. (2015) covers more than 300 languages, and includes parts of speech and morphological features.

Plank et al. (2015) discuss how various metrics for evaluating syntactic dependency parsing correlate with human judgments. We suggest that our translation-based metrics might naturally extend to dependency parsing by, e.g., treating an English dependency relation dictionary as a tag dictionary. The strong correlations between labeling (LA) and attachment scores (UAS) in dependency parsing favor our proposal.[5]

Garrette and Baldridge (2013) build taggers for low-resource languages from just 2 hours of manual annotation. Similarly, we show how to reliably evaluate cross-lingual POS taggers by translating as little as 100 most frequent English Wiktionary entries to the target language.

5 Conclusions

We evaluated how well the quality of POS taggers can be estimated *without annotated test data*. Our work has obvious applications to developing unsupervised or weakly supervised POS taggers for low-resource languages.

We were able to reliably estimate tagging accuracies by using very small tag dictionaries. Dictionaries with as little as 100 entries were in the majority of cases sufficient to predict true accuracies within 5%. We only require that these 100 entries be frequently used. Out of 5 competitive POS taggers, we then single out the best ones using our metric for 14/25 languages.

Finally, we showed that even if the dictionaries are "translated" from the English Wiktionary through a small list of bilingual word pairs we can still predict what POS taggers are best for 11/20 languages. In other words, we found that it is sufficient to translate a small list of frequent words from English to start reliably evaluating cross-lingual taggers for the true targets.

Acknowledgements

We acknowledge the three anonymous reviewers, and Natalie Schluter, for their valuable comments. Željko Agić and Barbara Plank thank the Nvidia Corporation for supporting their research. Anders Søgaard is funded by the ERC Starting Grant LOWLANDS No. 313695.

References

Željko Agić, Dirk Hovy, and Anders Søgaard. 2015. If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 268–272. Association for Computational Linguistics.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association of Computational Linguistics*, 4:301–312.

[4] `http://unimorph.org/`

[5] Pearson's $\rho = 0.82; 0.91$ (gold POS; predicted POS), UD data for 20 languages, TurboParser (Martins et al., 2013).

Thorsten Brants. 2000. TnT: A statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference*, pages 224–231.

Christos Christodouloupoulos and Mark Steedman. 2014. A massively parallel corpus: The Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375—-395.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 138–147. Association for Computational Linguistics.

William R. Knight. 1966. A computer method for calculating Kendall's tau with ungrouped data. *Journal of the American Statistical Association*, 61(314):436–439.

Taku Kudo. 2005. CRF++: Yet another CRF toolkit. `http://crfpp.sourceforge.net`.

Shen Li, João Graça, and Ben Taskar. 2012. Wiki-ly supervised part-of-speech tagging. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1389–1398. Association for Computational Linguistics.

Andre Martins, Miguel Almeida, and A. Noah Smith. 2013. Turning on the turbo: Fast third-order non-projective turbo parsers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 617–622. Association for Computational Linguistics.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97. Association for Computational Linguistics.

Thomas Mueller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332. Association for Computational Linguistics.

Joakim Nivre, Željko Agić, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Cristina Bosco, Sam Bowman, Giuseppe G. A. Celano, Miriam Connor, Marie-Catherine de Marneffe, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Daniel Galbraith, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Berta Gonzales, Bruno Guillaume, Jan Hajič, Dag Haug, Radu Ion, Elena Irimia, Anders Johannsen, Hiroshi Kanayama, Jenna Kanerva, Simon Krek, Veronika Laippala, Alessandro Lenci, Nikola Ljubešić, Teresa Lynn, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Shunsuke Mori, Hanna Nurmi, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Prokopis Prokopidis, Sampo Pyysalo, Loganathan Ramasamy, Rudolf Rosa, Shadi Saleh, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Kiril Simov, Aaron Smith, Jan Štěpánek, Alane Suhr, Zsolt Szántó, Takaaki Tanaka, Reut Tsarfaty, Sumire Uematsu, Larraitz Uria, Viktor Varga, Veronika Vincze, Zdeněk Žabokrtský, Daniel Zeman, and Hanzhi Zhu. 2015. Universal dependencies 1.2.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2089–2096. European Language Resources Association (ELRA).

Barbara Plank, Héctor Martínez Alonso, Željko Agić, Danijela Merkler, and Anders Søgaard. 2015. Do dependency parsing metrics correlate with human judgments? In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 315–320. Association for Computational Linguistics.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418. Association for Computational Linguistics.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015. A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 674–680. Association for Computational Linguistics.