One model, two languages: training bilingual parsers with harmonized treebanks

David Vilares, Miguel A. Alonso and Carlos Gómez-Rodríguez
Grypo LyS, Departamento de Computación, Universidade da Coruña
Campus de A Coruña s/n, 15071, A Coruña, Spain
{david.vilares, miguel.alonso, carlos.gomez}@udc.es

Abstract

We introduce an approach to train parsers using bilingual corpora obtained by merging harmonized treebanks of different languages, producing parsers that effectively analyze sentences in any of the learned languages, or even sentences that mix both languages. We test the approach on the Universal Dependency Treebanks, training with MaltParser and MaltOptimizer. The results show that these bilingual parsers are more than competitive, as some combinations not only preserve the performance, but even achieve significant improvements over the corresponding monolingual parsers. Preliminary experiments also show the approach to be promising on texts with code-switching.

1 Introduction

The need of frameworks for analysing content in different languages has been discussed recently (Dang et al., 2014), and multilingual dependency parsing is no stranger to this challenge. Data-driven parsing approaches (Nivre, 2006) provide models that can be trained for any given language, as long as enough annotated data is available.

On languages where treebanks are not available, cross-lingual transfer can be used to train parsers for a target language with data from one or more source languages. Data transfer approaches (e.g. Yarowsky et al. (2001), Tiedemann (2014)) map linguistic annotations across languages through parallel corpora. Instead, model transfer approaches (e.g. Naseem et al. (2012)) rely on cross-linguistic syntactic regularities to learn aspects of the source language that help parse an unseen language, without parallel corpora.

Model transfer approaches have benefitted from the development of multilingual resources that harmonize annotations across languages. Petrov et al. (2011) proposed a universal tagset of 12 PoS-tags, and McDonald et al. (2011) showed that it could be exploited to transfer delexicalized parsers (Zeman and Resnik, 2008). More recently, several projects have presented treebank collections of multiple languages with their annotations standardized at the syntactic level, including HamleDT (Zeman et al., 2012) and the Universal Dependency Treebanks (McDonald et al., 2013). The latter were used by Lynn et al. (2014) to train a set of delexicalized parsers for different languages, and evaluate them on an Irish treebank with the same annotation. They conclude that parser transfer can work across languages of different families under a common dependency annotation.

In this paper we also rely on those resources, but with a different purpose: we exploit universal annotations to train bilingual dependency parsers that effectively analyse unseen sentences in any of the learned languages. These bilingual parsers obtain surprisingly good results: our experiments show that, starting with a monolingual parsing, we can "teach" it an additional language for free in terms of accuracy (i.e., without significant accuracy loss in the original language, despite the increased complexity of the model to be learned) in the vast majority of cases. Furthermore, these bilingual parsers often even obtain significant improvements over the corresponding monolingual models on the original language. Preliminary experiments show that our bilingual parsers also work successfully on code-switching texts, where two languages can mix in the same sentence.

2 Bilingual training

Universal Dependency Treebanks v2.0 (McDonald et al., 2013) is a set of CoNLL-format treebanks for ten languages, annotated with common criteria. They include two versions of PoS tags: universal tags (Petrov et al., 2011) in the CPOSTAG column, and a refined annotation with language-specific information in the POSTAG column.

To train monolingual parsers (our baseline), we used the official training-dev-set splits provided with the corpora. For the bilingual models, for each pair of languages L_1, L_2; we simply merged their training sets into a single file acting as a training set for $L_1 \cup L_2$, and we did the same for the development sets. The test sets were not merged because comparing the bilingual parsers to monolingual ones requires evaluating each bilingual parser on the two corresponding monolingual test sets.

To build the models, we relied on MaltParser (Nivre et al., 2007). Due to the large number of language pairs that complicates manual optimization, and to ensure a fair comparison, we relied on MaltOptimizer (Ballesteros and Nivre, 2012), a system for the automatic optimization of MaltParser models. The software works in three phases: *Phase 1* and *2* choose a parsing algorithm by analyzing the training set, and performing experiments with default features. *Phase 3* tunes the feature model and learning algorithm parameters.

We propose two training configurations: (1) a *language-dependent tags* configuration where we include the information in the POSTAG column and (2) a *universal tags only* configuration, where we do not use this information, relying only on the CPOSTAG column. The aim is to find out if using universal or specific PoS tags has some impact on the performance of a bilingual parser. In this work, information that could be present in FEATS or LEMMA columns is not used. This methodology plans to answer two research questions: (1) can we train bilingual parsers with good accuracy by merging harmonized training sets?, and (2) is it essential that the tag sets for both languages are the same, or can we still get accuracy gains from fine-grained PoS tags (as in the monolingual case) even if they are language-specific?

3 Evaluation

To ensure a fair comparison between the monolingual and the bilingual models, we chose to optimize the models from scratch with MaltOptimizer, expecting it to choose the parsing algorithm and feature model which is more likely to obtain good results. Unexpectedly, we observed that the election of the bilingual parsing algorithm was not necessary related with the algorithms selected for the monolingual models. We observed that sometimes it chose one of the algorithms selected for any of the monolingual models and some others it chose a different parsing algorithm.

In view of this, and as it is known that different parsing algorithms can be more or less competitive depending on the language (Nivre, 2008), we ran a control experiment to evaluate the performance of the model setting the same parsing algorithm for all cases, executing only *phase 3* of MaltOptimizer. We chose the arc-eager parser for this experiment, as it was the algorithm that MaltOptimizer chose most frequently for the monolingual models in the previous configuration. The aim was to compare the accuracy of the bilingual models with respect to the monolingual ones, when there is no variation on the parsing algorithm between them. The results of this second experiment are not shown for space reasons, but they were very similar to those of the original experiment.

3.1 Results on the Universal Treebanks

Table 1 compares the performance of monolingual models with respect to bilingual ones, under the configuration *language-dependent tags*.

Each cell in the table contains the performance of a model, in terms of LAS and UAS. Cells in the diagonal correspond to monolingual models (the baseline), with the cell located at row i and column i representing the result obtained by training a monolingual parser on the training set of language L_i, and evaluating it on the test set of the same language L_i. Each cell outside the diagonal (at row i and column j, with $j \neq i$) contains the results of training a bilingual model on the training set for $L_i \cup L_j$, evaluated on the test set of L_i.

As we can see, in a large majority of the cases, bilingual parsers can learn to parse two languages with no statistically significant loss of accuracy with respect to the corresponding monolingual parsers ($p < 0,05$ with Bikel's randomized parsing evaluation comparator). This happened in 74 out of 90 cases when measuring UAS, or 69 out of 90 in terms of LAS. The implication of this is that, in most cases where we are applying a parser to sentences of a given language, adding a second language comes for free in terms of accuracy.

More surprisingly, there are many cases where bilingual parsers outperform monolingual ones, even in this evaluation on purely monolingual

R\ D	***de***	***en***	***es***	***fr***	***id***	***it***	***ja***	***ko***	***pt-br***	***sv***
de	78.27 84.03	78.01^{-} 84.08^{+}	77.82^{-} 83.82^{-}	77.83^{-} 83.55^{--}	77.84^{-} 83.85^{-}	78.10^{-} 84.12^{+}	77.86^{-} 83.88^{-}	77.94^{-} 83.63^{-}	78.13^{-} 83.87^{-}	78.60^{+} 84-38^{+}
en	89.37^{+} 91.02^{+}	89.36 91.02	89.46^{+} 91.09^{+}	89.38^{+} 91.06^{+}	89.69^{++} 91.32^{++}	89.82^{++} 91.47^{++}	89.43^{+} 91.10^{+}	89.63^{++} 91.32^{++}	89.60^{++} 91.24^{+}	89.11^{--} 90.79^{--}
es	80.85^{+} 85.17^{+}	81.08^{++} 85.27^{++}	80.60 84.75	80.95^{+} 85.15^{+}	81.16^{+} 85.00^{+}	80.92^{+} 85.13^{+}	81.41^{++} 85.52^{++}	81.49^{++} 85.39^{++}	79.96^{-} 84.70^{-}	81.26^{++} 85.42^{++}
fr	79.01^{-} 84.17^{-}	79.39^{+} 84.49^{+}	79.36^{+} 84.56^{+}	79.29 84.47	79.61^{+} 84.32^{-}	79.34^{+} 84.41^{-}	79.16^{-} 84.34^{-}	79.36^{+} 84.72^{+}	79.09^{-} 83.98^{-}	79.66^{+} 84.84^{+}
id	75.72^{--} 81.73^{--}	77.19^{-} 82.66^{--}	77.12^{-} 82.72^{-}	77.15^{-} 82.66^{-}	77.69 83.38	78.29^{+} 84.09^{+}	77.60^{-} 83.18^{-}	76.68^{--} 82.16^{--}	77.45^{-} 82.96^{-}	77.01^{--} 82.59^{--}
it	82.62^{--} 86.14^{--}	83.17^{--} 86.46^{--}	83.12^{--} 86.78^{-}	83.10^{--} 86.69^{-}	83.74^{--} 86.73^{--}	84.40 87.54	84.62^{+} 87.48^{-}	84.79^{+} 87.46^{-}	83.70^{-} 87.39^{-}	84.55^{+} 87.23^{-}
ja	76.53^{--} 83.77^{-}	76.24^{--} 83.89^{-}	76.61^{--} 84.26^{-}	76.32^{--} 84.05^{--}	75.18^{--} 83.08^{--}	77.05^{-} 83.97^{-}	77.46 84.34	76.89^{-} 83.65^{-}	76.69^{-} 83.97^{-}	76.89^{-} 84.17^{-}
ko	86.13^{--} 90.61^{--}	88.30^{+} 92.16^{+}	87.91^{+} 92.00^{-}	88.49^{+} 92.35^{+}	85.86^{--} 90.19^{--}	88.72^{++} 92.55^{+}	87.14^{--} 91.89^{-}	87.83 92.12	86.75^{--} 91.39^{--}	88.68^{-} 92.39^{-}
pt-br	84.83^{-} 87.18^{-}	85.06^{+} 87.19^{-}	84.99^{+} 87.27^{+}	84.97^{+} 87.17^{-}	85.10^{+} 87.35^{-}	85.43^{++} 87.68^{++}	84.95^{+} 87.13^{-}	85.12^{+} 87.35^{-}	84.88 87.39	85.25^{++} 87.43^{++}
sv	81.71^{--} 86.01^{--}	82.01^{--} 86.39^{-}	82.03^{-} 86.55^{-}	81.92^{--} 86.28^{--}	82.34^{-} 86.69^{-}	82.63^{+} 86.55^{-}	82.81^{+} 86.92^{+}	82.94^{++} 86.83^{-}	82.19^{-} 86.39^{-}	82.48 86.92

Table 1: Performance on the Universal Dependency Treebanks test sets using the POSTAG information. For each cell, its (row,column) pair indicates the language(s) with which the model was trained, with the row corresponding to the language where it was evaluated. '++' and '+' indicate that the improvement of performance obtained by the bilingual model is statistically significant or not, respectively. '- -' and '-' correspond to significant and not significant *decreases* in accuracy.

datasets. In particular, there are 12 cases where a bilingual parser obtains statistically significant gains in LAS over the monolingual baseline, and 9 cases with significant gains in UAS. As this clearly surpasses the amount of significant gains to be expected by chance, it is clear that there is synergy between datasets: in some cases, adding annotated data in a different language to our training set can actually improve the accuracy that we obtain in the *original* language. This opens up interesting research potential in using some confidence criterion to select the data that can help parsing in this way, akin to what is done in self-training approaches (Chen et al., 2008; Goutam and Ambati, 2011).

Looking at the results by language, we noted that the accuracy on the English and Spanish datasets almost always improves when adding a second treebank for training. Also interestingly, the only exceptions to this are when the English treebank is merged with the Swedish one, and Spanish with Portuguese, strongly suggesting (together with the rest of the results in the table) that merging treebanks of disparate languages can produce better results than closely-related languages, as in the transfer experiments by Lynn et al. (2014). Other languages that tend to get improvements by combining their treebanks with a second language are French and Portuguese. There seems to be a rough trend towards the languages with the largest training corpora benefiting from adding a second language, and those with the smallest corpora (like Indonesian, Italian or Japanese) suffering loss of accuracy, possibly because the training gets biased towards the second language.

Table 2 shows the performance of the monolingual and bilingual models under the *universal tags only* configuration. The bilingual parsers are also able to keep an acceptable performance with respect to the monolingual models, but significant accuracy losses are much more prevalent than under the *language-dependent tags* configuration.

This suggests that not only adding language-specific tagsets does not impair the training of bilingual models, but it is even benefitial. We hypothesize that this may be because the language-dependent PoS tags may help the parser identify specific constructions of one language that could cause confusion when parsing another one.

3.2 Parsing code-switched sentences

Apart from being able to parse sentences in different languages with a single model and without the need for prior language identification, an obvious application of our bilingual parsers is to analyze texts or utterances where both languages appear within the same sentence, i.e., sentences exhibiting code-switching. Unfortunately, there

R\ D	*de*	*en*	*es*	*fr*	*id*	*it*	*ja*	*ko*	*pt-br*	*sv*
de	74,07 79,77	72.04^{--} 77.52^{--}	74.51^{+} 79.95^{+}	74.44^{+} 79.83^{+}	73.68^{-} 79.24^{-}	73.76^{-} 79.44^{-}	73.90^{-} 79.83^{+}	74.30^{+} 79.76^{-}	74.29^{+} 79.71^{-}	74.76^{++} 80.25^{+}
en	88.46^{+} 90.35^{+}	88.35 90.27	88.65^{++} 90.54^{++}	88.39^{+} 90.26^{-}	88.61^{++} 90.47^{++}	88.68^{++} 90.53^{++}	88.65^{++} 90.49^{++}	88.61^{++} 90.43^{++}	88.65^{++} 90.55^{++}	88.50^{+} 90.43^{++}
es	79,66^{--} 83,81^{--}	78.78^{--} 82.94^{--}	80.54 84.35	79.59^{--} 83.26^{--}	78.98^{--} 82.79^{--}	79.84^{--} 83.79^{--}	79.59^{--} 83.53^{--}	79.80^{--} 83.57^{-}	79.74^{--} 83.76^{--}	79.09^{--} 83.28^{--}
fr	78,43^{+} 83,26^{-}	78.10^{-} 82.77^{-}	78.63^{+} 83.38^{-}	78.40 83.40	77.79^{-} 82.85^{-}	78.60^{+} 83.50^{+}	79.11^{+} 84.03^{+}	78.22^{-} 83.05^{-}	78,56^{+} 83,45^{+}	78.83^{+} 83.73^{+}
id	74.46^{--} 80.87^{--}	74.65^{--} 80.21^{--}	77.09^{--} 82.81^{--}	76.23^{--} 81.78^{--}	78.31 83.81	77.86^{-} 83.52^{-}	77.10^{--} 82.68^{--}	75.58^{--} 81.20^{--}	76.90^{--} 82.50^{--}	78.34^{+} 83.83^{+}
it	82.27^{--} 85.40^{--}	82.13^{--} 85.38^{--}	82.24^{--} 85.36^{--}	82.75^{--} 86.31^{--}	82.65^{--} 85.45^{--}	83.88 86.68	83.04^{--} 85.83^{--}	83.77^{-} 86.30^{-}	83.07^{--} 86.21^{--}	83.47^{-} 86.33^{-}
ja	69.41^{--} 79.62^{--}	68.88^{--} 79.21^{--}	69.28^{--} 79.45^{--}	69.24^{--} 80.11^{--}	69.73^{--} 79.58^{--}	70.22^{--} 79.58^{--}	70.87 81.16	69.73^{--} 80.23^{-}	69.24^{--} 79.37^{--}	70.02^{-} 80.47^{--}
ko	84.40^{--} 89.61^{--}	84.82^{--} 90.00^{--}	85.40^{--} 90.77^{--}	84.59^{--} 89.88^{--}	84.74^{--} 90.00^{--}	86.79^{-} 91.39^{-}	86.21^{--} 91.46^{--}	87.52 92.00	86.29^{--} 90.92^{--}	86.40^{--} 91.19^{--}
pt-br	83.40^{-} 85.78^{-}	82.76^{--} 85.01^{--}	83.56^{-} 85.82^{-}	83.72^{-} 85.85^{-}	83.08^{--} 85.38^{--}	83.95^{+} 86.15^{+}	83.80^{-} 85.93^{-}	84.16++ 86.33^{+}	83.83 86.11	84.28++ 86.41++
sv	79.65^{--} 84.14^{--}	79.61^{--} 84.42^{--}	79.75^{--} 84.46^{--}	80.46^{-} 84.88^{-}	80.94^{+} 85.14^{-}	81.06^{+} 85.51^{+}	81.19^{+} 85.29^{-}	81.11^{+} 85.14^{-}	80.89^{-} 85.05^{-}	80.93 85.32

Table 2: Performance on the Universal Dependency Treebanks test sets using the CPOSTAG information. The table is laid out with the same criteria as Table 1.

are no syntactically annotated code-switching corpora, so we could not perform a formal evaluation in this setting for lack of a gold standard.

We did perform informal tests, where we parsed some such sentences with the Spanish-English bilingual parsers. We observed that they were able to parse the English and Spanish part of the sentence much better than the monolingual models. Figure 1 illustrates an example.

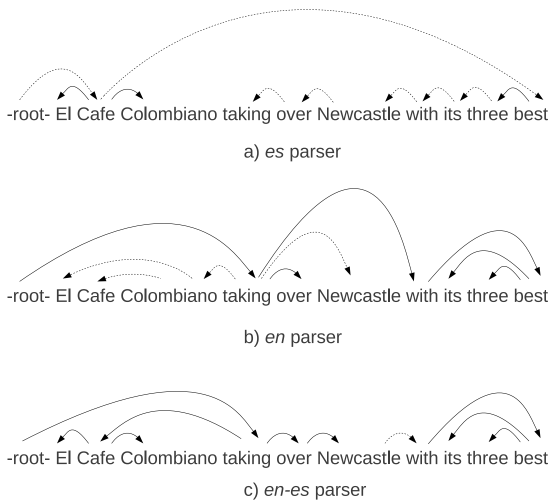

Figure 1: Example with the *en*, *es*, *en-es* models. Dotted lines represent incorrectly-parsed dependencies

This required training a bilingual tagger, which we did with the free distribution of the Stanford tagger (Toutanova and Manning, 2000); merging the Spanish and English corpora to train a combined bilingual tagger. Under the *universal tags only* configuration, the performance of the *en* and *es* taggers was 98.12%[1] and 96.03% respectively. The multilingual tagger obtained 98.00% and 95.88% over the monolingual test sets. Using language-dependent tags instead, the *en* and *es* taggers obtained 97.45% and 94.66% of accuracy. The multilingual tagger obtained 97,19% and 93.88% over the monolingual test sets.

4 Conclusions and future work

We have created bilingual parsers by merging corpora with a common annotation. Our results reflect that bilingual parsers do not lose accuracy with respect to monolingual parsers on their corresponding language, and can even outperform them, especially if fine-grained tags are used.

To our knowledge, this is the first attempt to train purely bilingual parsers that will analyze sentences irrespective of which of the two languages they are written in; as existing work on training a parsing model on two languages (Smith and Smith, 2004; Burkett and Klein, 2008) has focused on using parallel corpora to transfer linguistic knowledge between languages.

The applications include parsing sentences of

[1] Note that Toutanova and Manning reported 97,97% on the Penn Treebank tagset, which is bigger than the Google Universal tagset (48 vs 12 tags).

different languages with a single model, improving the accuracy of monolingual parsing with training sets from other languages, and successfully parsing sentences exhibiting code switching.

5 Acknowledgments

This research is supported by the Ministerio de Economía y Competitividad (FFI2014-51978-C2) and Xunta de Galicia (R2014/034). The first author is funded by the Ministerio de Educación, Cultura y Deporte (FPU13/01180).

References

[Ballesteros and Nivre2012] M. Ballesteros and J. Nivre. 2012. MaltOptimizer: an optimization tool for MaltParser. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 58–62. Association for Computational Linguistics.

[Burkett and Klein2008] D. Burkett and D. Klein. 2008. Two languages are better than one (for syntactic parsing). In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 877–886, Honolulu, Hawaii, October. Association for Computational Linguistics.

[Chen et al.2008] W. Chen, Y. Wu, and H. Isahara. 2008. Learning reliable information for dependency parsing adaptation. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 113–120, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Dang et al.2014] Y. Dang, Y. Zhang, Paul J. Hu, S. A. Brown, Y. Ku, J. Wang, and H. Chen. 2014. An integrated framework for analyzing multilingual content in Web 2.0 social media. *Decision Support Systems*, 61:126–135, May.

[Goutam and Ambati2011] R. Goutam and B. R. Ambati. 2011. Exploring self training for hindi dependency parsing. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1452–1456, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

[Lynn et al.2014] T. Lynn, J. Foster, M. Dras, and L. Tounsi. 2014. Cross-lingual Transfer Parsing for Low-Resourced Languages: An Irish Case Study. In *CLTW 2014. The First Celtic Language Technology Workshop. Proceedings of the Workshop*, pages 41–49, Dublin, Ireland, August.

[McDonald et al.2011] R. McDonald, S. Petrov, and K. Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 62–72. Association for Computational Linguistics.

[McDonald et al.2013] R. McDonald, J. Nivre, Y. Quirmbach-brundage, Y. Goldberg, D. Das, K. Ganchev, K. Hall, S. Petrov, Hao Zhang, O. Täckström, C. Bedini, N. Castelló, and J. Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 92–97. Association for Computational Linguistics.

[Naseem et al.2012] T. Naseem, R. Barzilay, and A. Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 629–637.

[Nivre et al.2007] J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, and E. Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

[Nivre2006] J. Nivre. 2006. Two strategies for text parsing. In Mickael Suominen, Antti Arppe, Anu Airola, Orvoki Heinämäki, Matti Miestamo, Urho Määttä, Jussi Niemi, Kari K. Pitkänen, and Kaius Sinnemäki, editors, *A Man of Measure: Festschrift in Honour of Fred Karlsson on his 60th Birthday*, pages 440–448. A special supplement to SKY Journal of Linguistics 19.

[Nivre2008] J. Nivre. 2008. Algorithms for Deterministic Incremental Dependency Parsing. *Computational Linguistics*, 34(4):513–553.

[Petrov et al.2011] S. Petrov, D. Das, and R. McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

[Smith and Smith2004] D. A. Smith and N. A. Smith. 2004. Bilingual parsing with factored estimation: Using english to parse korean. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 49–56, Barcelona, Spain, July. Association for Computational Linguistics.

[Tiedemann2014] J. Tiedemann. 2014. Rediscovering annotation projection for cross-lingual parser induction. In *Proceedings of COLING 2014: 25th International Conference on Computational Linguistics*, pages 1854–1864.

[Toutanova and Manning2000] K. Toutanova and C. D Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pages 63–70.

[Yarowsky et al.2001] D. Yarowsky, G. Ngai, and R. Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the 1st International Conference on Human Language Technology Research*, pages 1–8. Association for Computational Linguistics.

[Zeman and Resnik2008] D. Zeman and P. Resnik. 2008. Cross-language parser adaptation between related languages. In *IJCNLP 2008 Workshop on NLP for Less Privileged Languages*, pages 35–42, Hyderabad, India. International Institute of Information Technology.

[Zeman et al.2012] D. Zeman, D. Mareček, M. Popel, L. Ramasamy, J. Štěpánek, Z. Žabokrtský, and J. Hajič. 2012. HamleDT: To Parse or Not to Parse? 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. European Language Resources Association (ELRA).