Lexicalized vs. Delexicalized Parsing in Low-Resource Scenarios

Agnieszka Falenska and **Özlem Çetinoğlu**
Institute for Natural Language Processing
University of Stuttgart
{falenska,ozlem}@ims.uni-stuttgart.de

Abstract

We present a systematic analysis of lexicalized vs. delexicalized parsing in low-resource scenarios, and propose a methodology to choose one method over another under certain conditions. We create a set of simulation experiments on 41 languages and apply our findings to 9 low-resource languages. Experimental results show that our methodology chooses the best approach in 8 out of 9 cases.

1 Introduction

The recent CoNLL Shared Task on Parsing Universal Dependencies (CoNLL-ST) (Zeman et al., 2017) gave researchers the opportunity to study dependency parsing on a wide selection of treebanks. While the ultimate goal remained the same, i.e., achieving the best accuracy in predicting the head and dependency label of a token, the starting point changed from one group of languages to another, depending on the available resources.

In the *surprise languages* scenario, participants were given a very small training treebank, no development set, relatively accurate POS tags for the test set, and little or no parallel data.[1] When parallel data is not available, many of the standard cross-lingual parsing techniques (e.g. annotation projection (Hwa et al., 2005; McDonald et al., 2011); treebank translation (Tiedemann and Agić, 2016); utilizing cross-lingual word clusters (Täckström et al., 2012) or word embeddings (Duong et al., 2015; Guo et al., 2015)) become impossible to apply.

Delexicalized parsing (Zeman and Resnik, 2008) provides a suitable alternative, while it does not require parallel data. The central idea is to train a source-side parser without any lexical features, i.e., typically using only POS tags, and then use this trained parser to parse a target, low-resource language that shares the same POS tag set. No gold trees are required on the target side, and only POS tags have to be predicted prior to parsing. Given the simplicity of this method, several CoNLL-ST participants have chosen delexicalized approaches, not only for surprise languages but also for the other CoNLL-ST scenario – *small languages*[2]. In this scenario, as opposed to the surprise languages, the small training treebanks were the only source of gold POS tags.

When the data for training POS taggers is small – as for the small languages scenario as well as for many upcoming UD treebanks[3] – the delexicalized methods might be affected by poor POS accuracy. On the other hand, there are *some* gold trees for those scenarios and it is *possible* to train lexicalized parsers on them. Could there be cases in which such a low-resource lexicalized parser is preferred over a delexicalized one?

The central question we examine is whether we can find cases where a low-resource lexicalized parser achieves better accuracy than a delexicalized one. As a related problem, we investigate the following case: when there is a new language to parse, with no treebank but with the chance to predict POS tags, should one pursue a delexicalized parsing or invest in some tree annotation?

Our goal is to investigate those questions sys-

[1] This is due to the CoNLL-ST rules that restricted the use of parallel resources to OPUS (Tiedemann, 2012). But actually no or little parallel data is not an unrealistic assumption. Among four surprise languages, three of them have only Linux distro translations in OPUS, none are part of the 135-language Watchtower Corpus (Agić et al., 2016) (but two have a few documents on the Watchtower website), and none are part of the 100-language Edinburgh Bible Corpus (Christodouloupoulos and Steedman, 2015).

[2] The set of small treebanks with no development sets.

[3] For instance, currently there are 17 upcoming treebank projects within Universal Dependencies http://universaldependencies.org/#upcoming-ud-treebanks.

Proceedings of the 15th International Conference on Parsing Technologies, pages 18–24,
Pisa, Italy; September 20–22, 2017.

tematically in low-resource settings and to find the conditions under which one strategy leads to better results than the other. While our scenarios originate from the CoNLL-ST, our approach should be applicable to other settings. For example, our conclusions might prove helpful in developing early parsing models of a new treebank or in deciding how to proceed when there is a large gold POS tagged corpus but no trees (e.g. Echelmeyer et al. (2017) present a Middle High German corpus with 20,000 tokens of gold POS, no trees, and no apparent parallel data). They can also help plan resource creation. While POS annotation can be relatively fast (Garrette and Baldridge, 2013), creating treebanks is costly (Zeman and Resnik, 2008; Souček et al., 2013). The decision of building a large POS annotated corpus vs. a small treebank in a limited time, could depend on whether delexicalized models would work well for a target language.

2 Methodology

We compare low-resource lexicalized vs. delexicalized parsing in two settings used for surprise and small languages scenarios of the CoNLL-ST. In the first scenario, we assume the existence of a small treebank, and an external POS-annotated corpus – larger than the treebank – to train a POS tagger (**ExtPos**). In the second scenario, only the treebank exists, thus the gold POS tags necessary to train a tagger must be extracted from the treebank (**TBPos**). In both cases, we change the treebank sizes to observe the difference in accuracy between lexicalized and delexicalized parsing. While the POS accuracy is not affected from the treebank size in ExtPos, it changes with the size of the treebank in TBPos.

We employ multi-source delexicalized parsing (McDonald et al., 2011) in both scenarios. We follow Rosa and Žabokrtský (2015) and Agić (2017) in combining sources by blending (also known as reparsing) (Sagae and Lavie, 2006). Unlike in their original study, we apply blending on labeled arcs. We also employ weighted blending by assigning weights to sources based on language/treebank similarity measures.

2.1 Similarity Measures

We employ two measures of similarity between languages used in previous work:

KL$_{pos}$: Rosa and Žabokrtský's (2015) KL divergence metric (Kullback and Leibler, 1951) between POS trigrams. Instead of their smoothing, we use Laplace smoothing with $\alpha = 0.01$.

WALS: Agić's (2017) Hamming distance between each language's feature vectors from The World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013). For languages with no WALS entry, or for languages with less than 10 features in common, WALS is not defined.

2.2 Weights

We use three methods employing aforementioned similarity measures for weighting source languages while blending:

R&Z15: Rosa and Žabokrtský's (2015) KL$_{pos}$ calculation between the gold POS tags of the source and target training data, and their KL^{-4} weighting.

A17: Agić's (2017) combined weighting that calculates KL$_{pos}$ between the gold POS tags of the source training data and the predicted POS tags of the target development data, and then combines it with WALS.

LAS$_{tgt}$: We utilize gold trees as a source of weights for the blender. We parse the target training trees with the source delexicalized parsers and use their LAS as weights. Moreover, we rank the sources and take the n-best giving the best blended accuracy. n is tuned for every treebank and every training size separately on the training data.

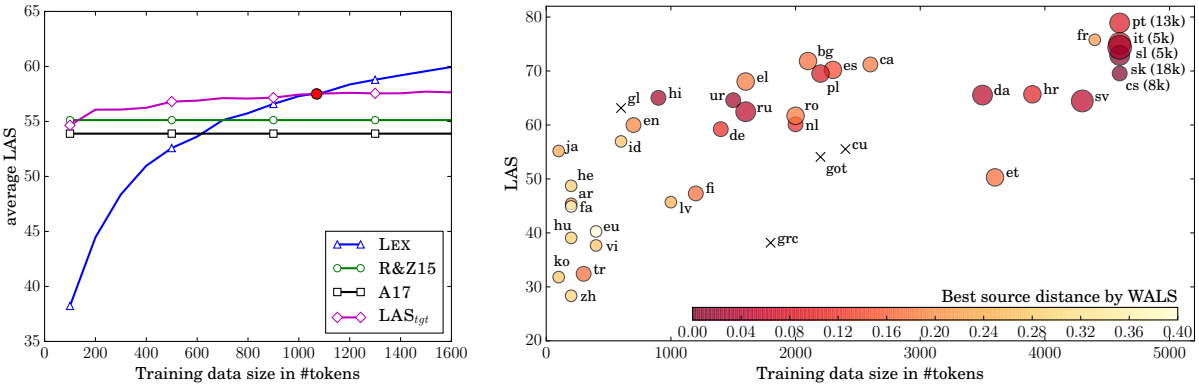
3 Experimental Setup

Data We use the Universal Dependencies v2.0 treebanks (Nivre et al., 2016) released for the CoNLL-ST (Nivre et al., 2017). We use all the treebanks except *domain-specific*[4] ones as sources (46 languages). As targets we take two groups of languages from the CoNLL-ST that correspond to the two settings we experiment with:

Surprise languages: Kurmanji (kmr), Upper Sorbian (hsb), North Sami (sme), Buryat (bxr). Each language contains a small sample of gold training data (see Table 1) and its test set is provided with POS tags predicted by a system trained on a data set much larger than the training data. Those languages represent the ExtPos setting.

Small languages: Latin (la), Irish (ga), Ukrainian (uk), Kazakh (kk), Uyghur (ug). They have small treebanks (especially Kazakh and

[4]Some languages have multiple UD treebanks, often from different domains. In such cases we chose the canonical treebank for a language.



(a) Average parsing accuracy. The average *cutting point* is marked by the red dot.

(b) Cutting points for all targets. Points which fall far right are moved left and their corresponding training sizes are given in brackets. cross (X): no WALS entry. circle size $\sim$ number of most similar sources (WALS < 0.2).

Figure 1: Parsing accuracies for EXTPOS scenario.

Uyghur), with no additional POS data. They correspond to the TBPOS setting.

The 9 target treebanks do not contain development sets, so as to not compromise our test sets, we set those languages aside as test cases. We perform analysis on simulated low-resource languages instead. For this purpose we use the subset of source treebanks that has a development set (41 languages). For each of them we sample small training treebanks starting with only 100 tokens.

Tools To train POS taggers, we employ MarMoT (Müller et al., 2013). In EXTPOS, POS taggers are trained on the whole treebanks, where in TBPOS, training is done only on small samples. We use Universal Part of Speech tags (UPOS) in all experiments.

For parsing, we use a beam-search transition-based parser (Björkelund and Nivre, 2015).[5] Delexicalized parsers (**DELEX**), and lexicalized parsers (**LEX**) for TBPOS are trained on gold POS tags. For EXTPOS lexicalized parsers are trained on 5-fold jackknifed POS tags for better performance (we sample the small treebanks after performing jackknifing). We blend delexicalized models' outputs via methods described in Section 2.2. In presenting the experimental results, we refer to these DELEX models with their weighting scheme, namely **R&Z15**, **A17**, and **LAS**$_{tgt}$.

Evaluation We use labeled attachment score (LAS) as the evaluation metric and evaluate using the script provided by the CoNLL-ST organizers.

[5]We also experimented with the graph-based parser Mate (Bohnet, 2010) to test our hypotheses on a different parsing architecture. We achieved results parallel to the transition-based parser, thus we only present one set of results.

4 Results and Analysis

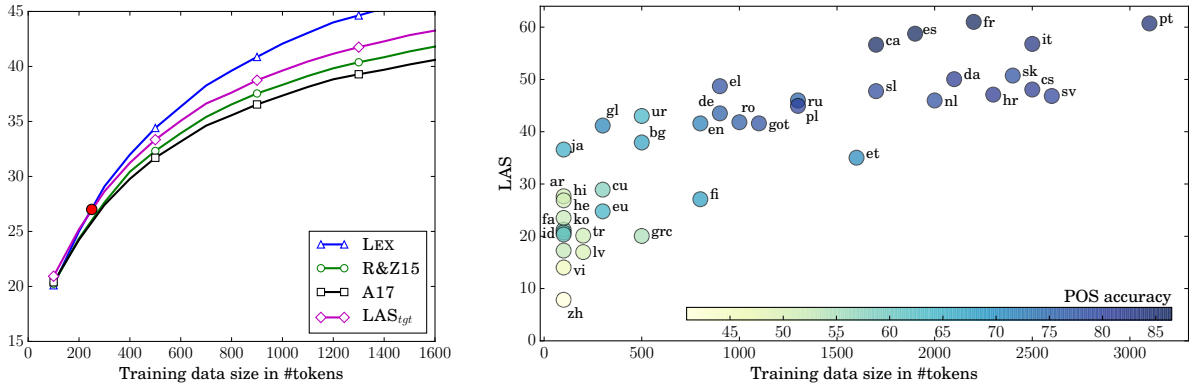
We apply LEX and DELEX to our artificially created low-resource languages and analyze the results for EXTPOS and TBPOS scenarios.

4.1 The EXTPOS Scenario

Figure 1(a) shows average results for EXTPOS. We observe that DELEX methods using POS tags as weight source (A17 and R&Z15) achieve on average 55% LAS. Both of them are surpassed by LEX when gold trees are available for only 750 tokens (which corresponds to 40 sentences for the used treebanks). Creating weights from gold trees (LAS$_{tgt}$) instead of POS tags improves the performance but the improvement is modest and LEX trained on only 1,100 tokens (avg. 55 sentences) outperforms it. We find this small number surprising, and investigate it further.

Instead of looking at averages we analyze target languages separately. We call the point in which LEX surpasses LAS$_{tgt}$ (the red dot in Figure 1(a)) a *cutting point* and plot the cutting points for all the target languages separately (Figure 1(b)). We note that most of the languages do not fall around the average cutting point of 1,100 tokens. Instead they can be grouped into three clusters: 10 languages on the far left for which LEX trained on even less than 500 tokens (avg. 25 sentences) is better than DELEX; 9 languages on the far right for which even 3,400 tokens (avg. 176 sentences) is not enough to surpass DELEX; and the middle part. In order to explain this distribution we take a look at the target languages characteristics.

We use WALS to measure the distance between languages (we normalize WALS by the number of



(a) Average parsing accuracy. The average *cutting point* is marked by the red dot.

(b) Cutting points for all target languages.

Figure 2: Parsing accuracies for TBPOS scenario.

common features). For every target we select the best source according to WALS and represent its distance by color (the darker the circle the more similar the best source is). The number of good sources (WALS < 0.2) is represented by size (the bigger the circle the more good sources exist). For example, Korean's (ko) best source according to WALS is Urdu with a distance of 0.27 and its circle is light and small. In contrast, Slovenian (sl) for which five good sources exist (among which Ukrainian has the smallest distance of 0.03) is represented by a dark and big circle.

In Figure 1(b) we can observe a pattern among the two border groups. The left group tends to have small and light circles which means that there is no good source for them. When we look at the languages which fall into this group (like Arabic (ar) and Vietnamese (vi)) we see that they come from language families less represented among the source languages. The far right circles in comparison are bigger and darker which means that they fit well into the set of existing source languages. Indeed many Slavic languages fall here.

4.2 The TBPOS Scenario

Figure 2(a) shows average results for TBPOS. As expected the accuracy of parsers drops due to lower POS accuracy. Other than the LAS scores, there are two major differences between those two plots. In this setting, POS taggers are trained on tags coming from the treebanks and POS accuracy changes with the data size. That is why the A17 and R&Z15 lines are not flat any more. The difference between them is once again very slight and LAS$_{tgt}$ outperforms both of them for all data sizes.

As the second major difference to EXTPOS, LEX quickly outperforms DELEX methods at around 250 tokens. We show the breakdown of this cutting point by target language in Figure 2(b). The shades of the circles denote the POS accuracy (the darker the more accurate). In this case, the circles are placed in a smaller area with an upper limit of 3,100 tokens. Note that the source language similarity is not shown in this plot, yet it still affects the underlying distribution of targets. Comparing to Figure 1(b), the relative placement of the languages is almost the same. The dense distribution also causes a more homogeneous scatter making the groupings less visible.

4.3 Overall Picture

Delexicalized models use only POS tags as features and therefore are more influenced by low POS tagging accuracy than lexicalized parsers. That is why the choice which method to apply depends strongly on the existing resources.

In EXTPOS in order for DELEX to work better than LEX good sources must exist. If the target belongs to an underrepresented language family, even a very small sample of gold trees is enough for LEX to achieve better results. In contrast, if the target is similar to many sources DELEX can help even if the target gold data exists. In that case, the gold trees can be exploited for example as a source for weights for blending. For all our targets regardless of existing sources, having a treebank bigger than 18,000 tokens is enough for LEX to give higher accuracies than any DELEX model.

In TBPOS on average LEX outperforms DELEX even when trained on only few gold sentences. The only situation when using other sources might help is when there are many similar sources and

gold trees for more than 1,000 tokens (avg. 50 sentences). In that case, the POS tagging accuracy is able to reach a reasonable 70% and DELEX achieves comparable performance to LEX. But already having a treebank bigger than only 3,100 tokens (avg. 160 sentences) is enough for LEX to work better for all our targets.

5 Application to Test Languages

We use findings from Section 4 to decide which test languages should employ DELEX methods.

In EXTPOS all the treebanks have less than 500 tokens and according to Figure 1(a) DELEX methods might be a good choice. We compare the test languages with the breakdown in Figure 1(b). Buryat is a Mongolic language and does not have any close relatives among the source languages. For Kurmanji there is only one similar language – Persian. Therefore for both of those languages the cutting point would likely occur quickly and DELEX would give no or very little improvement over LEX. On the contrary, North Sami is Finno-Ugric and Finnish and Estonian should be good sources for it. Upper Sorbian is Slavic and its family is well represented (e.g. by Czech or Polish). Therefore DELEX should work very well for them.

In TBPOS, most of the languages are much bigger than 3,100 tokens and for none of them DELEX should help. Kazakh is smaller than the threshold of 1,000 tokens. Most probably POS tagging accuracy for it would be poor and LEX should be a better choice to overcome that. The only language for which DELEX methods might help is Uyghur. But it is a language for which not many good sources exist, the closest being Turkish. Most probably LEX is also a better choice in this case.

To test our hypotheses, we apply all the methods to the test languages and present results in Table 1. For languages not present in WALS (Kazakh, Uyghur) A17 uses only KL_{pos}. We do not apply R&Z15 to the surprise languages since their POS tagging training data is not available[6]. We see that our intuition was right in 8 out of 9 cases. For all the small languages LEX performs better. For the surprise languages as expected Upper Sorbian and North Sami gain from DELEX the most – 9.22 and 6 points LAS respectively when compared to LAS_{tgt}. For Kurmanji LEX trained on only 242 tokens (20 sentences) gives 2.12 points

[6]Their test sets were annotated via jackknifing by the CoNLL-ST organizers to mimic EXTPOS scenario.

		size	POS	LEX	R&Z15	A17	LAS_{tgt}
Small	la	18184	84.41	**41.02**	33.62	35.06	35.06
	ga	13826	89.99	**64.66**	41.75	42.17	44.54
	uk	12846	88.58	**64.81**	62.20	58.46	63.67
	ug	1662	74.96	**34.81**	21.04	20.72	30.11
	kk	529	58.48	**26.55**	21.49	24.49	26.17
Surprise	hsb	460	90.30	49.59	–	57.09	**58.81**
	kmr	242	90.04	**40.94**	–	38.82	38.17
	bxr	153	84.12	28.06	–	29.07	**32.01**
	sme	147	86.81	29.8	–	33.44	**35.80**

Table 1: Results on the test languages.

LAS more than the best DELEX. Surprisingly, for Buryat both DELEX methods outperform LEX.

6 Conclusion

In this paper we looked into lexicalized vs. delexicalized parsing in cases where there are few trees for targets, POS tagging accuracies for the test set vary, and no reasonable amount of parallel data between source and target languages is available. To systematically compare these two approaches and to observe under what circumstances one is more favorable than the other, we created a simulation scenario of 41 low-resource languages and applied our findings to a set of 9 real low-resource targets.

We found out that lexicalized parsing can surpass delexicalized methods even when trained on very few sentences, so one should not be deceived by the small size of a target treebank. By analyzing the typological relations between the source and target languages, and the accuracy of POS tagging it is possible to develop intuition about which of two methods to apply to a new language. For 8 out of 9 test languages our findings hold true.

In our experiments we assumed specific constraints on existing resources, e.g., no parallel data between source and target languages. If some parallel data is available, other transfer parsing approaches should be taken into comparison. For instance, Agić et al. (2016) analyze a related scenario where parallel data is available but no gold trees. They show that projecting POS and dependency annotations from multiple source languages outperforms single-best delexicalized parsing as well as blending. Whether such a method would be a preferable choice when a small amount of gold trees is available is a venue to explore.

Acknowledgments

This project is funded by the Deutsche Forschungsgemeinschaft (DFG) via the SFB 732, projects D2 and D8 (PI: Jonas Kuhn).

References

Željko Agić. 2017. Cross-lingual parser selection for low-resource languages. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies*. Gothenburg Sweden, pages 1–10.

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

Anders Björkelund and Joakim Nivre. 2015. Non-deterministic oracles for unrestricted non-projective transition-based dependency parsing. In *Proceedings of the 14th International Conference on Parsing Technologies*. Association for Computational Linguistics, Bilbao, Spain, pages 76–86.

Bernd Bohnet. 2010. Very high accuracy and fast dependency parsing is not a contradiction. In *Proc. of COLING*.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation* 49(2):375.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig. http://wals.info/.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. A neural network model for low-resource universal dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 339–348.

Nora Echelmeyer, Nils Reiter, and Sarah Schulz. 2017. Ein PoS-Tagger fur das Mittelhochdeutsche . In *Book of Abstracts of DHd 2017* . Bern, Switzerland. https://doi.org/10.18419/opus-9023.

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*. Association for Computational Linguistics, Atlanta, Georgia, pages 138–147.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2015. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of ACL-IJCNLP*. Beijing, China, pages 1234–1244.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural language engineering* 11(3):311–325.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* pages 79–86.

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, pages 62–72.

Thomas Müller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of EMNLP*. Seattle, Washington, USA, pages 322–332.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Cristina Bosco, Gosse Bouma, Sam Bowman, Marie Candito, Gülşen Cebirolu Eryiit, Giuseppe G. A. Celano, Fabricio Chalub, Jinho Choi, Çar Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökrmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Linh Hà M, Dag Haug, Barbora Hladká, Petter Hohle, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşkara, Hiroshi Kanayama, Jenna Kanerva, Natalia Kotsyba, Simon Krek, Veronika Laippala, Phng Lê Hng, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Lng Nguyn Th, Huyn Nguyn Th Minh, Vitaly Nikolaev, Hanna Nurmi, Stina Ojala, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalnia, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Livy Real, Laura Rituma, Rudolf Rosa, Shadi Saleh, Manuela Sanguinetti, Baiba Saulīte, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jonathan North

Washington, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 2017. Universal dependencies 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University. http://hdl.handle.net/11234/1-1983.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Rudolf Rosa and Zdeněk Žabokrtský. 2015. Klcpos3 - a language similarity measure for delexicalized parser transfer. In *Proceedings of ACL-IJCNLP*.

Kenji Sagae and Alon Lavie. 2006. Parser combination by reparsing. In *Proceedings of NAACL*. pages 129–132.

Milan Souček, Timo Järvinen, and Adam LaMontagne. 2013. Managing a multilingual treebank project. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*. Charles University in Prague, Matfyzpress, Prague, Czech Republic, Prague, Czech Republic, pages 292–297. http://www.aclweb.org/anthology/W13-3732.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montreal, Canada, NAACL HLT '12, pages 477–487.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey.

Jörg Tiedemann and Željko Agić. 2016. Synthetic treebanking for cross-lingual dependency parsing. *Journal of AI Research* 55(1):209–248.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinkova, Jan Hajic jr., Jaroslava Hlavacova, Václava Kettnerová, Zdenka Uresova, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria dePaiva, Kira Droganova, Héctor Martínez Alonso, Çağr Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonca, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics, Vancouver, Canada, pages 1–19.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*. Hyderabad, India.