KL_{cpos^3} – a Language Similarity Measure for Delexicalized Parser Transfer

Rudolf Rosa and **Zdeněk Žabokrtský**
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, Prague, Czech Republic
{rosa, zabokrtsky}@ufal.mff.cuni.cz

Abstract

We present KL_{cpos^3}, a language similarity measure based on Kullback-Leibler divergence of coarse part-of-speech tag trigram distributions in tagged corpora. It has been designed for multilingual delexicalized parsing, both for source treebank selection in single-source parser transfer, and for source treebank weighting in multi-source transfer. In the selection task, KL_{cpos^3} identifies the best source treebank in 8 out of 18 cases. In the weighting task, it brings +4.5% UAS absolute, compared to unweighted parse tree combination.

1 Introduction

The approach of delexicalized dependency parser transfer is to train a parser on a treebank for a source language (src), using only non-lexical features, most notably part-of-speech (POS) tags, and to apply that parser to POS-tagged sentences of a target language (tgt) to obtain dependency parse trees. Delexicalized transfer yields worse results than a supervised lexicalized parser trained on a target language treebank. However, for languages with no treebanks available, it may be useful to obtain at least a lower-quality parse tree for tasks such as information retrieval.

Usually, multiple source treebanks are available, and it is non-trivial to select the best one for a given target language. As a solution, we present a language similarity measure based on KL divergence (Kullback and Leibler, 1951) of distributions of coarse POS tag trigrams in POS-tagged corpora, which we call KL_{cpos^3}. The measure has been designed and tuned specifically for multilingual delexicalized parser transfer, and it often succeeds in selecting the best source treebank in a single-source setting, as well as in appropriately weighting the source treebanks by similarity to the target language in a multi-source parse tree combination approach.

2 Related Work

Delexicalized parser transfer was conceived by Zeman and Resnik (2008), who also introduced two important preprocessing steps – mapping treebank-specific POS tagsets to a common set using Interset (Zeman, 2008), and harmonizing treebank annotation styles into a common style, which later developed into the HamleDT harmonized treebank collection (Zeman et al., 2012).

McDonald et al. (2011) applied delexicalized transfer in a setting with multiple source treebanks available, finding that the problem of selecting the best source treebank without access to a target language treebank for evaluation is non-trivial. They combined all source treebanks by concatenating them but noted that this yields worse results than using only the best source treebank.

An alternative is the (monolingual) parse tree combination method of Sagae and Lavie (2006), who apply several independent parsers to the input sentence and combine the resulting parse trees using a maximum spanning tree algorithm. Surdeanu and Manning (2010) enrich tree combination with weighting, assigning each parser a weight based on its Unlabelled Attachment Score (UAS). In our work, we introduce an extension of this method to a crosslingual setting by combining parsers for different languages and using source-target language similarity to weight them.

Several authors (Naseem et al., 2012; Søgaard and Wulff, 2012; Täckström et al., 2013b) employed WALS (Dryer and Haspelmath, 2013) to estimate source-target language similarity for delexicalized transfer, focusing on genealogy distance and word-order features. Søgaard and Wulff (2012) also introduced weighting into the treebank concatenation approach, using a POS n-gram model trained on a target-language corpus

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 243–249,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

to weight source sentences in a weighted perceptron learning scenario (Cavallanti et al., 2010). KL divergence (Kullback and Leibler, 1951) of POS tag distributions, as well as several other measures, was used by Plank and Van Noord (2011) to estimate monolingual domain similarity.

As is quite common in parsing papers, including those dealing with semi-supervised and unsupervised parsing, we use gold POS tags in all our experiments. This enables us to evaluate the effectiveness of our parsing method alone, not influenced by errors stemming from the POS tagging. Based on the published results, it seems to be considerably easier to induce POS tags than syntactic structure for under-resourced languages, as there are several high-performance weakly-supervised POS taggers. Das and Petrov (2011) report an average accuracy of 83% using word-aligned texts, compared to 97% reached by a supervised tagger. Täckström et al. (2013a) further improve this to 89% by leveraging Wiktionary. For some languages, there are even less resources available; Agić et al. (2015b) were able to reach accuracies around 70% by using partial or full Bible translation. Our methods could thus be applied even in a more realistic scenario, where gold POS tags are not available for the target text, by using a weakly-supervised POS tagger. We intend to evaluate the performance of our approach in such a setting in future.

3 Delexicalized Parser Transfer

Throughout this work, we use MSTperl (Rosa, 2015b), an implementation of the unlabelled single-best MSTParser of McDonald et al. (2005b), with first-order features and non-projective parsing, trained using 3 iterations of MIRA (Crammer and Singer, 2003).[1]

Our delexicalized feature set is based on the set of McDonald et al. (2005a) with lexical features removed. It consists of combinations of signed edge length (distance of head and parent, bucketed for values above 4 and for values above 10) with POS tag of the head, dependent, their neighbours, and all nodes between them.[2] We use the Universal POS Tagset (UPOS) of Petrov et al. (2012).

3.1 Single-source Delexicalized Transfer

In the single-source parser transfer, the delexicalized parser is trained on a single source treebank, and applied to the target corpus. The problem thus reduces to selecting a source treebank that will lead to a high performance on the target language.

3.2 Multi-source Delexicalized Transfer

In our work, we extend the monolingual parse tree combination method to a multi-source crosslingual delexicalized parser transfer setting:

1. Train a delexicalized parser on each source treebank.
2. Apply each of the parsers to the target sentence, obtaining a set of parse trees.
3. Construct a weighted directed graph as a complete graph over all tokens of the target sentence, where each edge is assigned a score equal to the number of parse trees in which it appears (each parse tree contributes by either 0 or 1 to the edge score). In the weighted variant of the method, the contribution of each parse tree is multiplied by its weight.
4. Find the final dependency parse tree as the maximum spanning tree over the graph, using the algorithm of Chu and Liu (1965) and Edmonds (1967).

4 KL_{cpos^3} Language Similarity

We introduce KL_{cpos^3}, a language similarity measure based on distributions of coarse POS tags in source and target POS-tagged corpora. This is motivated by the fact that POS tags constitute a key feature for delexicalized parsing.

The distributions are estimated as frequencies of UPOS trigrams[3] in the treebank training sections:

$$f(cpos_{i-1}, cpos_i, cpos_{i+1}) = \\ = \frac{\text{count}(cpos_{i-1}, cpos_i, cpos_{i+1})}{\sum_{\forall cpos_{a,b,c}} \text{count}(cpos_a, cpos_b, cpos_c)}; \quad (1)$$

we use a special value for $cpos_{i-1}$ or $cpos_{i+1}$ if $cpos_i$ appears at sentence beginning or end.

We then apply the Kullback-Leibler divergence

[1] Note that while our approach does not depend in principle on the actual parser used, our results and conclusions may not be valid for other parsers.

[2] The feature set, as well as scripts and configuration files for the presented experiments, are available in (Rosa, 2015a).

[3] Bigrams and tetragrams performed comparably on the weighting task, but worse on the selection task. Using more fine-grained POS tags led to worse results as fine-grained features tend to be less shared across languages.

$D_{\mathrm{KL}}(tgt||src)$ to compute language similarity:[4]

$$\begin{aligned} &KL_{cpos^3}(tgt, src) = \\ &= \sum_{\forall cpos^3 \in tgt} f_{tgt}(cpos^3) \cdot \log \frac{f_{tgt}(cpos^3)}{f_{src}(cpos^3)}, \qquad (2) \end{aligned}$$

where $cpos^3$ is a coarse POS tag trigram. For the KL divergence to be well-defined, we set the source count of each unseen trigram to 1.

4.1 KL_{cpos^3} for Source Selection

For the single-source parser transfer, we compute KL_{cpos^3} distance of the target corpus to each of the source treebanks and choose the closest source treebank to use for the transfer.

4.2 $KL^{-4}_{cpos^3}$ for Source Weighting

To convert KL_{cpos^3} from a negative measure of language similarity to a positive source parser weight for the multi-source tree combination method, we take the fourth power of its inverted value.[5] The parse tree produced by each source parser is then weighted by $KL^{-4}_{cpos^3}(tgt, src)$.

5 Dataset

We carry out our experiments using HamleDT 2.0 of Rosa et al. (2014), a collection of 30 treebanks converted into Universal Stanford Dependencies (de Marneffe et al., 2014), with POS tags converted into UPOS; we use gold-standard POS tags in all experiments. We use the treebank training sections for parser training and language similarity estimation, and the test sections for evaluation.[6]

5.1 Tuning

To avoid overfitting the exact definition of KL_{cpos^3} and $KL^{-4}_{cpos^3}$ to the 30 treebanks, we used only 12 *development treebanks* for hyperparameter tuning: ar, bg, ca, el, es, et, fa, fi, hi, hu, it, ja.[7]

Measure	Avg	SD	Best
$KL^{-4}_{cpos^3}(tgt, src)$	**51.0**	**16.7**	**6**
$KL^{-4}_{cpos^3}(src, tgt)$	50.6	17.4	4
$JS^{-4}_{cpos^3}(tgt, src)$	49.6	18.0	2
$cos_{cpos^3}(tgt, src)$	49.0	17.7	1

Table 1: Weighted multi-source transfer using various similarity measures. Evaluation using average UAS on the development set.

Avg = Average UAS.
SD = Standard sample deviation of UAS, serving as an indication of robustness of the measure.
Best = Number of targets for which the measure scored best.

Table 1 contains evaluation of several language similarity measures considered in the tuning phase, applied to weighted multi-source transfer and evaluated using average UAS on the development set. We evaluated KL divergences computed in both directions, as well as Jenses-Shannon divergence (Lee, 2001) and cosine similarity. Based on the results, $KL^{-4}_{cpos^3}$ was selected, as it performed best in all aspects.

Once the hyperparameters were fixed, we applied the parser transfer methods to the full set of 30 treebanks; our final evaluation is based on the results on the 18 test treebanks as targets.

5.2 Other datasets

Additionally, we also report preliminary results on the Prague style conversion of HamleDT, which loosely follows the style of the Prague Dependency Treebank of Böhmová et al. (2003), and on the subset of CoNLL 2006 and 2007 shared tasks (Buchholz and Marsi, 2006; Nilsson et al., 2007) that was used by McDonald et al. (2011).[8]

6 Evaluation

6.1 Results

Table 2 contains the results of our methods both on the test languages and the development languages.

[4] The KL divergence is non-symmetric; $D_{\mathrm{KL}}(P||Q)$ expresses the amount of information lost when a distribution Q is used to approximate the true distribution P. Thus, in our setting, we use $D_{\mathrm{KL}}(tgt||src)$, as we try to minimize the loss of using a *src* parser as an approximation of a *tgt* parser.

[5] A high value of the exponent strongly promotes the most similar source language, giving minimal power to the other languages, which is good if there is a *very* similar source language. A low value enables combining information from a larger number of source languages. We chose a compromise value of 4 based on performance on the development data.

[6] Contrary to the motivation, we do not evaluate our method on truly underresourced languages, since automatic intrinsic evaluation is not possible on languages without treebanks. Still, e.g., Bengali and Telugu can be considered low-resourced, since their treebanks are very small.

[7] We tuned the choice of the similarity measure, POS n-gram length, and the way of turning KL_{cpos^3} into $KL^{-4}_{cpos^3}$. To tune our method to perform well in many different situations, we chose the development set to contain both smaller and larger treebanks, a pair of very close languages (ca, es), a very solitary language (ja), multiple members of several language families (Uralic, Romance), and both primarily left-branching (bg, el) and right-branching (ar, ja) languages.

[8] The CoNLL subset is: da, de, el, en, es, it, nl, pt, sv.

For each target language, we used all remaining 29 source languages for training (in the single-source method, only one of them is selected and applied).

Our baseline is the treebank concatenation method of McDonald et al. (2011), i.e., a single delexicalized parser trained on the concatenation of the 29 source treebanks.

As an upper bound,[9] we report the results of the oracle single-source delexicalized transfer: for each target language, the oracle source parser is the one that achieves the highest UAS on the target treebank test section.[10] For space reasons, we do not include results of a higher upper bound of a supervised delexicalized parser (trained on the target treebank), which has an average UAS of 68.5%. It was not surpassed by our methods for any target language, although it was reached for Telugu, and approached within 5% for Czech and Latin.

6.2 Discussion

The results show that KL_{cpos^3} performs well both in the selection task and in the weighting task, as both the single-source and the weighted multi-source transfer methods outperform the unweighted tree combination on average, as well as the treebank concatenation baseline. In 8 of 18 cases, KL_{cpos^3} is able to correctly identify the oracle source treebank for the single-source approach. In two of these cases, weighted tree combination further improves upon the result of the single-source transfer, i.e., surpasses the oracle; in the remaining 6 cases, it performs identically to the single-source method. This proves KL_{cpos^3} to be a successful language similarity measure for delexicalized parser transfer, and the weighted multi-source transfer to be a better performing approach than the single-source transfer.

The weighted tree combination is better than its unweighted variant only for half of the target languages, but it is more stable, as indicated by its lower standard deviation, and achieves an average UAS higher by 4.5% absolute. The unweighted tree combination, as well as treebank concatenation, perform especially poorly for English, German, Tamil, and Turkish, which are rich in determiners, unlike the rest of the treebanks;[11] there-

Tgt	TB	Oracle		Single-src			Multi-src	
lang	conc	del trans		KL			$\times 1$	$\times w$
bn	61.0	te	**66.7**	0.5	**te**	**66.7**	63.2	**66.7**
cs	60.5	sk	**65.8**	0.3	**sk**	**65.8**	60.4	**65.8**
da	**56.2**	en	**55.4**	0.5	sl	42.1	**54.4**	50.3
de	12.6	en	**56.8**	0.7	**en**	**56.8**	27.6	**56.8**
en	12.3	de	**42.6**	0.8	**de**	**42.6**	21.1	**42.6**
eu	**41.2**	da	**42.1**	0.7	tr	29.1	**40.8**	30.6
grc	43.2	et	42.2	1.0	sl	34.0	**44.7**	42.6
la	38.1	grc	40.3	1.2	cs	35.0	**40.3**	39.7
nl	55.0	da	57.9	0.7	**da**	57.9	56.2	**58.7**
pt	62.8	en	64.2	0.2	es	62.7	**67.2**	62.7
ro	44.2	it	**66.4**	1.6	la	30.8	**51.2**	50.0
ru	55.5	sk	57.7	0.9	la	40.4	**57.8**	57.2
sk	52.2	cs	**61.7**	0.2	sl	58.4	**59.6**	58.4
sl	45.9	sk	**53.9**	0.2	**sk**	**53.9**	47.1	**53.9**
sv	45.4	de	**61.6**	0.6	da	49.8	**52.3**	50.8
ta	27.9	hi	**53.5**	1.1	tr	31.1	28.0	**40.0**
te	67.8	bn	**77.4**	0.4	**bn**	**77.4**	68.7	**77.4**
tr	18.8	ta	40.3	0.7	**ta**	40.3	23.2	**41.1**
Test	44.5		**55.9**	0.7		48.6	48.0	**52.5**
SD	16.9		10.8			14.4	15.0	11.8
ar	37.0	ro	**43.1**	1.7	sk	41.2	35.3	**41.3**
bg	64.4	sk	66.8	0.4	**sk**	66.8	66.0	**67.4**
ca	56.3	es	**72.4**	0.1	**es**	**72.4**	61.5	**72.4**
el	63.1	sk	61.4	0.7	cs	60.7	62.3	**63.8**
es	59.9	ca	**72.7**	0.0	**ca**	**72.7**	64.3	**72.7**
et	67.5	hu	71.8	0.9	da	64.9	70.5	**72.0**
fa	30.9	ar	**35.6**	1.1	cs	**34.7**	32.5	33.3
fi	41.9	et	44.2	1.1	**et**	44.2	41.7	**47.1**
hi	24.1	ta	**56.3**	1.1	fa	20.8	24.6	**27.2**
hu	55.1	et	52.0	0.7	cs	46.0	**56.5**	51.2
it	52.5	ca	**59.8**	0.3	pt	54.9	59.5	**59.6**
ja	29.2	tr	**49.2**	2.2	ta	**44.9**	28.8	34.1
Dev	48.5		**57.1**	0.9		52.0	50.3	**53.5**
SD	15.2		12.5			16.1	16.5	16.7
All	46.1		**56.4**	0.8		50.0	48.9	**52.9**
SD	16.1		11.3			15.0	15.4	13.7
PRG test			**60.0**			49.7	55.7	**58.1**
PRG dev			**64.0**			57.5	58.0	**61.1**
PRG all			**61.5**			52.8	56.6	**59.3**
CoNLL			**58.3**			53.1	**58.1**	55.7

Table 2: Evaluation using UAS on test target treebanks (upper part of the table) and development target treebanks (lower part).

For each target language, all 29 remaining non-target treebanks were used for training the parsers. The best score among our transfer methods is marked in bold; the baseline and upper bound scores are marked in bold if equal to or higher than that.

Legend:
Tgt lang = Target treebank language.
TB conc = Treebank concatenation.
Oracle del trans = Single-source delexicalized transfer using the oracle source language.
Single-src = Single-source delexicalized transfer using source language with lowest KL_{cpos^3} distance to the target language (language bold if identical to oracle).
Multi-src = Multi-source delexicalized transfer, unweighted ($\times 1$) and $KL_{cpos^3}^{-4}$ weighted ($\times w$).
Test, *Dev*, *All*, *SD* = Average on test/development/all, and its standard sample deviation.
PRG, *CoNLL* = Preliminary results (average UAS) on Prague conversion of HamleDT, and on subset of CoNLL used by McDonald et al. (2011).

[9] This is a hard upper-bound for the single-source transfer, but can be surpassed by the multi-source transfer.

[10] We do not report the matrix of all source/target combination results, as this amounts to 870 numbers.

[11] In the treebanks for these four languages, determiners constitute around 5-10% of all tokens, while most other treebanks contain no determiners at all; in some cases, this is

fore, determiners are parsed rather randomly.[12] In the weighted methods, this is not the case anymore, as for a determiner-rich target language, determiner-rich source languages are given a high weight.

For target languages for which KL_{cpos^3} of the closest source language was lower or equal to its average value of 0.7, the oracle treebank was identified in 7 cases out of 12 and a different but competitive one in 2 cases; when higher than 0.7, an appropriate treebank was only chosen in 1 case out of 6. When KL_{cpos^3} failed to identify the oracle, weighted tree combination was always better or equal to single-source transfer but mostly worse than unweighted tree combination. This shows that for distant languages, KL_{cpos^3} does not perform as good as for close languages.

We believe that taking multiple characteristics of the languages into account would improve the results on distant languages. A good approach might be to use an empirical measure, such as KL_{cpos^3}, combined with supervised information from other sources, such as WALS. Alternatively, a backoff approach, i.e. combining KL_{cpos^3} with e.g. KL_{cpos^2}, might help to tackle the issue.

Still, for target languages dissimilar to any source language, a better similarity measure will not help much, as even the oracle results are usually poor. More fine-grained resource combination methods are probably needed there, such as selectively ignoring word order, or using different sets of weights based on POS of the dependent node.

6.3 Evaluation on Other Datasets

In (Rosa, 2015c), we show that the accuracies obtained when parsing HamleDT treebanks in the Universal Stanford Dependencies annotation style are significantly lower than when using the Prague style. Preliminary experiments using the Prague style conversion of HamleDT generally show our methods to be effective even on that dataset, although the performance of KL_{cpos^3} is lower in source selection – it achieves lower UAS than unweighted tree combination, and only identifies the oracle source treebank in 30% cases. This may be due to us having used only the Stanfordized treebanks for tuning the exact definition of the measure.

Preliminary trials on the subset of CoNLL used by McDonald et al. (2011) indicated that our methods do not perform well on this dataset. The best results by far are achieved by the unweighted combination, i.e., it is best not to use KL_{cpos^3} at all on this dataset. We believe this to be a deficiency of the dataset rather than of our methods – it is rather small, and there is low diversity in the languages involved, most of them being either Germanic or Romanic. The HamleDT dataset is larger and more diverse, and we believe it to correspond better to the real-life motivation for our methods, thus providing a more trustworthy evaluation.

In the near future, we intend to reevaluate our methods using the Universal Dependencies treebank collection (Nivre et al., 2015; Agić et al., 2015a), which currently contains 18 languages of various types and seems to be steadily growing. A potential benefit of this collection is the fact that the annotation style harmonization seems to be done with more care and in a more principled way than in HamleDT, presumably leading to a higher quality of the dataset.

7 Conclusion

We presented KL_{cpos^3}, an efficient language similarity measure designed for delexicalized dependency parser transfer. We evaluated it on a large set of treebanks, and showed that it performs well in selecting the source treebank for single-source transfer, as well as in weighting the source treebanks in multi-source parse tree combination.

Our method achieves good results when applied to similar languages, but its performance drops for distant languages. In future, we plan to explore combinations of KL_{cpos^3} with other language similarity measures, so that similarity of distant languages is estimated more reliably.

In this work, we only used the unlabelled first-order MSTParser. We intend to also employ other parsers in future, possibly in combination, and in a labelled as well as unlabelled setting.

Acknowledgments

This research was supported by the grants GAUK 1572314, SVV 260 224, and FP7-ICT-2013-10-610516 (QTLeap). This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

related to properties of the treebank annotation or its harmonization rather than properties of the language.

[12] UAS of determiner attachment tends to be lower than 5%, which is several times less than for any other POS.

References

Željko Agić, Maria Jesus Aranzabe, Aitziber Atutxa, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richárd Farkas, Jennifer Foster, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Jan Hajič, Anders Trærup Johannsen, Jenna Kanerva, Juha Kuokkala, Veronika Laippala, Alessandro Lenci, Krister Lindén, Nikola Ljubešić, Teresa Lynn, Christopher Manning, Héctor Alonso Martínez, Ryan McDonald, Anna Missilä, Simonetta Montemagni, Joakim Nivre, Hanna Nurmi, Petya Osenova, Slav Petrov, Jussi Piitulainen, Barbara Plank, Prokopis Prokopidis, Sampo Pyysalo, Wolfgang Seeker, Mojgan Seraji, Natalia Silveira, Maria Simi, Kiril Simov, Aaron Smith, Reut Tsarfaty, Veronika Vincze, and Daniel Zeman. 2015a. Universal Dependencies 1.1. `http://hdl.handle.net/11234/LRT-1478`.

Željko Agić, Dirk Hovy, and Anders Søgaard. 2015b. If all you have is a bit of the bible: Learning POS taggers for truly low-resource languages. In *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015)*. Hrvatska znanstvena bibliografija i MZOS-Svibor.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2003. The Prague dependency treebank. In *Treebanks*, pages 103–127. Springer.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 149–164. Association for Computational Linguistics.

Giovanni Cavallanti, Nicolo Cesa-Bianchi, and Claudio Gentile. 2010. Linear algorithms for online multitask classification. *The Journal of Machine Learning Research*, 11:2901–2934.

Yoeng-Jin Chu and Tseng-Hong Liu. 1965. On shortest arborescence of a directed graph. *Scientia Sinica*, 14(10):1396.

Koby Crammer and Yoram Singer. 2003. Ultraconservative online algorithms for multiclass problems. *The Journal of Machine Learning Research*, 3:951–991.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 600–609. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Natalia Silveira, Timothy Dozat, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proc. of LREC'14*, Reykjavík, Iceland. ELRA.

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

Jack Edmonds. 1967. Optimum branchings. *Journal of Research of the National Bureau of Standards B*, 71(4):233–240.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, pages 79–86.

Lillian Lee. 2001. On the effectiveness of the skew divergence for statistical language analysis. In *Artificial intelligence and statistics*, volume 2001, pages 65–72.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005a. Online large-margin training of dependency parsers. In *Proceedings of the 43rd annual meeting on ACL*, pages 91–98. ACL.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005b. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530. ACL.

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*, EMNLP '11, pages 62–72, Stroudsburg, PA, USA. ACL.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the ACL: Long Papers - Volume 1*, ACL '12, pages 629–637, Stroudsburg, PA, USA. ACL.

Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL shared task session of EMNLP-CoNLL*, pages 915–932. sn.

Joakim Nivre, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richárd Farkas, Jennifer Foster, Filip Ginter, Yoav Goldberg, Jan Hajič, Jenna Kanerva, Veronika Laippala, Alessandro Lenci, Teresa Lynn, Christopher Manning, Ryan McDonald, Anna Missilä, Simonetta Montemagni, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Maria Simi, Aaron Smith, Reut Tsarfaty, Veronika Vincze, and Daniel Zeman. 2015. Universal Dependencies 1.0. `http://hdl.handle.net/11234/1-1464`.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proc. of LREC-2012*, pages 2089–2096, Istanbul, Turkey. ELRA.

Barbara Plank and Gertjan Van Noord. 2011. Effective measures of domain similarity for parsing. In *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies-Volume 1*, pages 1566–1576. ACL.

Rudolf Rosa, Jan Mašek, David Mareček, Martin Popel, Daniel Zeman, and Zdeněk Žabokrtský. 2014. HamleDT 2.0: Thirty dependency treebanks Stanfordized. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2334–2341, Reykjavík, Iceland. ELRA.

Rudolf Rosa. 2015a. MSTperl delexicalized parser transfer scripts and configuration files. `http://hdl.handle.net/11234/1-1485`.

Rudolf Rosa. 2015b. MSTperl parser (2015-05-19). `http://hdl.handle.net/11234/1-1480`.

Rudolf Rosa. 2015c. Multi-source cross-lingual delexicalized parser transfer: Prague or Stanford? In Eva Hajičová and Joakim Nivre, editors, *Proceedings of the Third International Conference on Dependency Linguistics, Depling 2015*, Uppsala, Sweden. Uppsala University.

Kenji Sagae and Alon Lavie. 2006. Parser combination by reparsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 129–132. ACL.

Anders Søgaard and Julie Wulff. 2012. An empirical study of non-lexical extensions to delexicalized transfer. In *COLING (Posters)*, pages 1181–1190.

Mihai Surdeanu and Christopher D. Manning. 2010. Ensemble models for dependency parsing: Cheap and good? In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, HLT '10, pages 649–652, Stroudsburg, PA, USA. ACL.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013a. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013b. Target language adaptation of discriminative transfer parsers. In *Proceedings of NAACL-HLT 2013*, pages 1061–1071.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *IJCNLP 2008 Workshop on NLP for Less Privileged Languages*, pages 35–42, Hyderabad, India. Asian Federation of Natural Language Processing, International Institute of Information Technology.

Daniel Zeman, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. 2012. HamleDT: To parse or not to parse? In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. ELRA.

Daniel Zeman. 2008. Reusable tagset conversion using tagset drivers. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 213–218, Marrakech, Morocco. ELRA.