What is hard in Universal Dependency Parsing?

Angelika Kirilin[a] and **Yannick Versley**[a,b]
[a]Institute for Computational Linguistics, [b]Leibniz ScienceCampus
University of Heidelberg
(kirilin|versley)@cl.uni-heidelberg.de

Abstract

Verifying (or at least attempting to falsify) claims about one language being more hard to parse than another, or about one parser being applicable to a maximally wide range of languages, used to quickly devolve into apples-and-oranges comparison since different languages, i.e., different treebanks also mean different annotation schemes and a different source of text.

In this paper, we use two datasets that contain annotations for several languages according to Universal Stanford Dependencies (UniDepTB 1.0 and HamleDT 2.0) to perform a comparison of different parsers, languages and (partially) annotation schemes both at a coarse level and at a finer level of characterizing differences by typical patterns of the errors in context.

1 Introduction

Analyzing and understanding the performance of a parser is crucial for determining where additional complexity or effort can help, and also in order to understand where model combination(s) can or cannot help. At the very coarsest level, one might consider the results from a single evaluation metric (e.g. UAS, LAS, ParsEval, ...), and try to find explanations or correlates in different factors that may affect the difficulty of the task, especially treebank size and language, but possibly also annotation scheme.

Beyond such coarse-grained approaches looking at the overall quality of parser output, once can also look at at individual parses (e.g. using the WhatsWrong visualization tool[1]). For languages one is not intimately familiar with, or when the types of errors are relatively diverse, this approach will drown the user in too much information.

[1] code.google.com/whatswrong

This is why, in this paper, we attempt to find a middle ground between these two extremes by going from overall measures to phenomenon-specific measures, to substructures setting errors regarding these phenomena in context, all while trying to correlate the phenomena, and the relevant substructures, to influencing factors such as parsers, annotation schemes, and treebank size.

1.1 Finer-grained Evaluation

McDonald and Nivre (2007) analyze the results from the CoNLL-X shared task on dependency parsing in multiple languages. Because McDonald and Nivre's work predate universal dependency schemes, they have to manually identify related dependency labels across annotation schemes and assume that attachment difficulty is comparable across annotation schemes. Still, McDonald and Nivre are able to show that the accuracy of the dependencies produced shows that certain dependencies (as characterized by the part-of-speech of the dependent) are more difficult than others, including those of adpositions and conjunctions.

Kummerfeld et al. (2012) compare various parsers on the Penn Treebank regarding several specific constructions, including the attachment of prepositional phrases, temporal noun phrases, clauses as well as coordinations, NP-internal structures and several others, showing that in certain cases (adding self-training, adding a reranker to the Charniak parser, looking at oracle parses for larger and larger k-best lists) a uniform effect on all constructions can be found, whereas differences between varying genres of the Brown corpus used as out-of-domain testing set, or between very different parsing models show more pronounced differences between different categories.

1.2 Mining for Idiosyncrasies

Goldberg and Elhlad (2010) analyze different parsers by training a tree boosting classifier to dis-

Proceedings of the 6th Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2015), pages 31–38,
Bilbao, Spain, July 23rd 2015.

tinguish between test set sentences and sentences from parser output, with the intuition that constructions that a parser *overproduces* are taken by the classifier as indications for sentences from the parser, whereas those constructions that are usually *underproduced* are good indicators for sentences that came from the gold-annotated test set.

In an approach to find idiosyncratic structures in dependency trees, Dickinson (2010) looks at rule expansions (i.e. one head and its dependent) as well as children bigrams that are rare or not found in the training corpus, and shows that it is possible to find erroneous dependencies with some success.

1.3 Comparing Ingredients

Schwartz et al. (2012) use the Penn Treebank in conjunction with a customized version of the LTH dependency converter of Johansson and Nugues (2007) to investigate how several parsers perform on variant conversions of the treebank. For several phenomena, they found that parsing results strongly preferred one particular alternative: in coordination, using the first conjunct as a head is strongly preferred over using the conjunction as a head; in noun phrases, using the noun as a head is strongly preferred over using the determiner, and in prepositional phrases, using the preposition as a head is preferred over using the noun phrase's head instead. They also found a (less pronounced) preference for modals and "*to*" particles as the heads of complex verb phrases.

The work of Popel et al. (2013) focuses more narrowly on coordination structures in treebanks covering 26 languages, and perform roundtrip experiments for a conversion to annotation in the style of the Prague treebank and back to the original structures, noting that usually, but not always, the roundtrip is possible with very little in information loss.

Comparing results within different annotation schemes is also the motivation for the work of Tsarfaty et al. (2011), who propose to evaluate parses across schemes by creating a generalized tree that contains the information common to two annotation schemes.

In the remainder of **this paper**, we will first give a brief overview over the treebanks and parsers used (§2), discuss our approach to link overall quality measures and coarse-grained measures for certain phenomena to influencing factors (§3.1), finishing by drilling down into the most typical contexts in which this error-ingredient interaction plays out (§3.2).

2 Materials

2.1 Treebanks

The **HamleDT** set of treebanks (Zeman et al., 2012) contains different dependency treebanks converted to the annotation style of the Prague Dependency Treebank (Böhmova et al., 2001), in particular using conjunctions as the head in coordination, and using long part-of-speech tags with multiple feature slots instead of using a separate field for morphological properties. The conversion tool used in HamleDT has been extended to produce a version in Stanford dependencies (Rosa et al., 2014). In the Stanfordized versions of HamleDT, we have the short part-of-speech tags of Petrov et al. (2012), a longer version containing additional information, and an additional layer of morphological properties.

The **Universal Dependencies Treebank** 1.0 (McDonald et al., 2013) contains only trees matching the Stanford Dependencies schema, consisting in part of trees converted from existing treebanks, and in part from newly annotated text. The version 1 of the Universal Treebank only contains the short part-of-speech tags of Petrov et al. (2012).[2]

We used five languages that are at the intersection of those covered by both HamleDT and the Universal Dependencies Treebank, namely *German*, *English*, *Finnish*, *Swedish* and *Spanish*. While for some languages, the textual material is the same (the Finnish treebank is the same between both schemes, while the German part of HamleDT is entirely from the Tiger treebank and the corresponding part of UniDepTB contains some sentences from Tiger but also some newly annotated sentences with social media text).

2.2 Parsers

MaltParser (Hall et al., 2006) is a transition-based parser for several different transition systems (arc-eager, arc-standard, using either a pseudoprojective transform or a swapping transition to account for nonprojective dependencies. To find sensible settings for transition system, features, and SVM hyperparameters, we used the MaltOptimizer software of Ballesteros and Nivre (2012).

[2]The newer Version 2.0 of the Universal Dependencies treebank does contain morphological tags in addition to the coarse part-of-speech tags, but uses a different file format from that used in HamleDT and UniDepTB 1.0.

The **MATE parser** uses a transition-based system with beam search and a scoring system based on third-order factors (Bohnet and Kuhn, 2012), which makes it competitive to other third-order graph-based parser despite using a transition-based search graph to form its hypotheses. The MATE parser is able to reach nonprojective parses using hill-climbing from a projective solution.

TurboParser (Martins et al., 2013) uses a dual decomposition approach for the decoding of dependency trees with third-order factors.

The **RBG parser** (Zhang et al., 2014) is a third-order parser with additional global features that we use in two variants: the first, which we reference as *RBG1* in the following text, uses exact decoding on first-order edges using the Chu-Liu-Edmonds algorithm as proposed by McDonald et al. (2005). The second, which uses the full capabilities of the RBG parser and which we call *RBG3* in the following text, uses hill-climbing with random restarts to find a good solution for a model using third-order factors and additional global features.

3 Experiments

In the following, we perform an analysis of parsing experiments that have been done for the five different parsers (*Malt+Optimizer*, *MATE*, *Turbo*, *RBG1*, *RBG3*) across five different languages (*German*, *English*, *Finnish*, *Swedish*, *Spanish*) using two variants of the Universal Dependency Treebank (normal plus content-head) plus five variations of HamleDT with PDT or Stanfordized trees and short, long and (for the Stanfordized version) long plus morphology tags. Certain combinations of languages and schemes/tagsets (e.g., Finnish with standard Universal Dependencies, English with content-head Universal Dependencies) were not available, which means that we cover most, but not all of these combinations.

In addition to the cross-product of parsers/languages/schemes for all datasets, we added parsing results for *RBG1* and *RBG3* for all the languages in standardized training set sizes of 3.500 sentences and (where available) 10.000 sentences.

3.1 Factorial Analysis

In order to detect general trends among the results of the experiments we performed with many influencing factors (essentially a design that has combinations of different languages, parsering models and annotation schemes). Because, e.g., Finnish *also* has a rather small treebank, uses the content-head variant of Stanford dependencies and no morphological tags, a regression model can potentially give a clearer picture than just looking at averages in this case.

We fitted a linear model predicting several measures of success such as labeled accuracy (LAS), accuracy for coordination (`Coord` in PDT scheme, `cc` in the USD scheme), for subjects and objects (`Sb`/`nsubj` and `Obj`/`obj`,`iobj`,`dobj`). We fit a linear model including the original factors as well as a limited number of interactions that were selected using Akaike's Information Criterion (AIC), which provides a tradeoff between (LAS-prediction) likelihood and number of model parameters. We used the Ordinary Least Squares implementation of the `statmodels` package.

Looking at table 1, and the subset of experiments reported in table 2, we find some very general findings: having more data is beneficial, the content-head schema seems to make things significantly worse. We see that German and Finnish (languages with free(r) word order) profit most from long POS tags and morphological information, while English and Swedish do not, and that RBG1 and RBG3 results seem to benefit strongly from the short POS tags.[3]

Looking at the parser-size and language-size interaction, we see two large outliers – Finnish gaining over 5 LAS points for every doubling in training size, and Swedish only about 0.25 points with each doubling. These outliers may have occurred due to the fact that our experiments do not contain data points for larger treebank sizes.

3.2 Specific Phenomena

In order to be able to reason about specifics in the behaviour of parsers in multiple settings, or for multiple languages, we used the gSpan algorithm (Yan and Han, 2002) to extract patterns with a minimum frequency as well as minimum and maximum size.

We can then use weighting functions to rank patterns by some interesting properties, or to compare their frequency across two different conditions.

[3]This can be confirmed by a look into the parser code, which contains special cases for the universal tags of conjunctions, adpositions, punctuation and verbs.

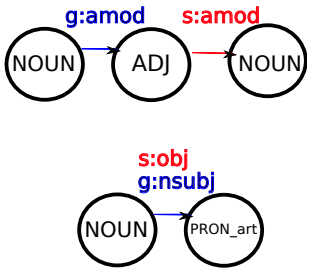


Figure 1: Two of the most characteristic patterns (by χ^2) for Spanish

Frequent Error patterns One question we have is whether there are kinds of errors that are typical of particular parsers, annotation schemes, or other variables. To assess this, we construct an acyclic graph formed by forming the (discriminated) union of system and gold edges (marking edges that are system-only, gold-only, or labeled differently). Using gSpan, we filtered for those patterns that were (a) frequent enough, and (b) contained an edge that indicated non-matching of gold tree and system output.

For selecting patterns based on statistical significance, we computed expected counts (based on error patterns mined from all experiments) and observed counts (error pattern frequency in the condition that we focus on), and compute Pearson's χ^2 statistic to rank subgraphs that are characteristic for (i.e., strongly associated with) a particular subset of the graphs, for example in Figure 1 contrasting the most typical error patterns for German with those for Spanish in the parses of the RBG1 parser. In that particular instance, we discover some idiosyncracies of HamleDT's conversion to universal dependencies of the German or Spanish side; the German conversion contains sentence dependencies as a governor (presumably in gapping or coordination), whereas Spanish has a `PRON_art` POS tag that the other languages do not share, and the possibility of attaching (some) adjectives either to the right and to the left.

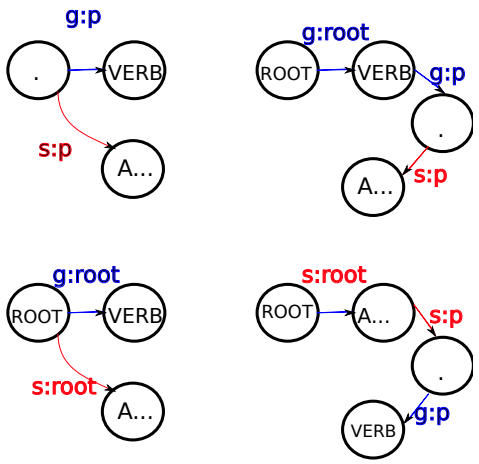
Overproduced structures To get an idea if today's parser have a bias for over- or underproducing certain structures, we counted for every frequent-enough structure how often it occurs in the gold standard, and in the output of a given parser, respectively.

Using frequent error patterns to compare RBG1 and RBG3 on all subgraphs that occur at least 20 times, for example, we find that the edge-factored RBG1 model has a strong tendency to over- and underproduce certain structures (overproducing 36% and underproducing 22% of pattern *types*) whereas RBG3 slightly reduces these type counts (with 40% and 20% of these subgraphs, respectively), with a relatively stronger decrease in the actual occurrences of these patterns. Like Goldberg and Elhadad, we find that the idea of over- and underproduction finds a very substantial number of such structures for simpler models (such as MaltParser's deterministic shift-reduce model or the edge-factored RBG1) while this tendency is much diminished for the state-of-the-art third-order models.

Small treebank versus crosslingual parsers In recent research, approaches to port parsers to languages without treebanks using either projection or model transfer have received a lot of attention – justifiably so, since they could help save significant effort in building treebanks.

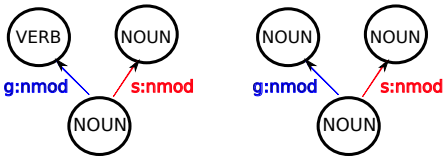
Here we compare the output of the state-of-the-art crosslingual parser of Guo et al. (2015), which uses word clusters and embeddings to generalize over words in different languages, with a first-order factored model (RBG1) trained on 3 500 sentences (which would admittedly still take a large amount of time to produce, but is less than 10% of the size of the larger treebanks in German, English or Czech).

Maybe surprisingly, the errors that the crosslingual models makes more frequently with a large margin (in between 19% and 20% of all sentences) all concern root edges and punctuation:

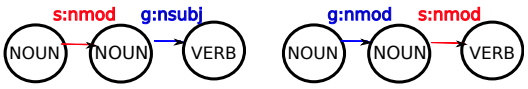


Larger treebank versus smaller treebank Considering the annotation of data, we may also ask what we gain from annotating a treebank that

is significantly larger, but not extremely so (i.e., 10 000 sentences rather than 3 500 sentences). In this case, many of the errors that the larger-treebank parser can avoid (between 1.6% and 1.8% of sentences) are cases of PP attachment or genitives (USD label `nmod`).



Weaker versus stronger parsing model Similarly to going from a weaker to a stronger parsing model, we find that the patterns with the largest normalized frequency difference are those involving PP attachment; due to our method of filtering (requiring both a false positive and a false negative edge), we find a number of patterns that show correlated errors (with differences between 0.8% and 1.3% of all sentences):



4 Summary

In this paper, we addressed the question of the influence of languages, parsing models and part-of-speech representations in universal dependency parsing in two ways, one on a very coarse level (aggregating over different experiments with the same or roughly comparable annotation schemes), and one on a finer but still aggregate level (mining for error patterns that are considerably more frequent in one setting than in another).

Our first proposal, aggregating parsing results not by averaging but in a regression model, is preferable to the former because it allows us to try to separate out multiple influences – for example, Finnish generally being a difficult language for today's parsers because of its morphological structure *as well as* having a relatively small treebank. While the model we present here is still relatively simple, we hope it will inspire more complex models that can predict parsing accuracy based on a more diverse set of factors (e.g., morphological richness, counts for unknown words, or sentence length).

The subgraph mining approach in the second part of the paper has the advantage over older approaches such as Goldberg's that it can represent more complex structures, including those that contain both system and gold-standard parses as in our error pattern approach. The use of normalized frequency differences instead of using a boosting-based linear classifier has both the advantage and the disadvantage of yielding higher weights for structures with many variants where a regularized linear classifier would either shrink all weights to a smaller size (L2 regularization) or shrink the weights for all but one representative for a group of several related patterns (L1 regularization).

In this paper, we have addressed the question of assessing the influence of languages and annotation schemes in universal dependency parsing, a question which had only partially been addressed before.

We show that when using gold part-of-speech tags, the improvements from going from a smaller treebank to a larger one, or from a weaker (edge-factored) parser to a stronger (third-order/global) one are dominated by PP attachment effects; however it is not clear whether these effects would persist with realistic morphosyntactic tagging (as POS and morphology errors would often propagate to syntax), nor whether using semi-supervised parsing techniques (which can easily improve PP attachment) would result in a different distribution of errors.

Acknowledgements The authors thank the anonymous reviewers for their insightful comments. Many thanks to Jiang Guo *et al.*, who kindly provided the output of their crosslingual parser.

References

Ballesteros, M. and Nivre, J. (2012). MaltOptimizer: A system for MaltParser optimization. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2012)*.

Böhmova, A., Hajič, J., Hajičová, E., and Hladká, B. (2001). The Prague dependency treebank: Three-level annotaion scenario. In *Treebanks: Building and using syntactically annotated corpora*, pages 103–127. Kluwer Academic Publishers.

Bohnet, B. and Kuhn, J. (2012). The best of both worlds - a graph-based completion model for transition-based parsers. In *EACL 2012*.

Dickinson, M. (2010). Detecting errors in automatically-parsed dependency relations. In *ACL 2010*.

Goldberg, Y. and Elhlad, M. (2010). Inspecting the structural biases of dependency parsing algorithms. In *CoNLL 2010*.

Guo, J., Che, W., Yarowsky, D., Wang, H., and Liu, T. (2015). Cross-lingual dependency parsing based on distributed representations. In *Proceedings of ACL 2015*.

Hall, J., Nivre, J., and Nilsson, J. (2006). Discriminative classifiers for deterministic dependency parsing. In *Coling-ACL 2006*.

Johansson, R. and Nugues, P. (2007). Extended constituent-to-dependency conversion for english. In *NODALIDA 2007*.

Kummerfeld, J. K., Hall, D., Curran, J. R., and Klein, D. (2012). Parser showdown at the wall street corral: An empirical investigation of error types in parser output. In *EMNLP 2012*.

Martins, A., Almeida, M., and Smith, N. (2013). Turning on the turbo: Fast third-oreder non-projective turbo parsers. In *ACL 2013*.

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

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of ACL 2013*, pages 92–97.

McDonald, R., Pereira, F., Ribarov, K., and Hajič, J. (2005). Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of HLT/EMNLP 2005*.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*.

Popel, M., Marecek, D., Stepanek, J., Zeman, D., and Zabortsky, Z. (2013). Coordination structures in dependency treebanks. In *ACL 2013*.

Rosa, R., Masek, J., Marecek, D., Popel, M., Zeman, D., and Zabortsky, Z. (2014). Hamledt 2.0: Thirty dependency treebanks stanfordized. In *LREC 2014*.

Schwartz, R., Abend, O., and Rappoport, A. (2012). Learnability-based syntactic annotation design. In *Coling 2012*.

Tsarfaty, R., Nivre, J., and Andersson, E. (2011). Evaluating dependency parsing: Robust and heuristics-free cross-annotation evaluation. In *EMNLP 2011*.

Yan, X. and Han, J. (2002). gSpan: Graph-based substructure pattern mining. In *Proceedings fo the Second IEEE Conference on Data Mining (ICDM 2002)*.

Zeman, D., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., and Hajič, J. (2012). Hamledt: To parse or not to parse? In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Zhang, Y., Lei, T., Barzilay, R., and Jaakkola, T. (2014). Greed is good if randomized: New inference for dependency parsing. In *EMNLP 2014*.

Appendix A: Factor Analysis (all schemes/parsers)

Property		Average LAS	Δ Avg LAS	OLS Loadings LAS	Coord	Subj	Obj
Global Average/Intercept		82.10	—	72.90	48.72	73.56	61.19
size		—	—	1.30	3.12	2.06	2.29
Parsers							
MaltParser		76.63	-5.46	—	—	—	—
MaltOptimizer		80.83	-1.27	3.70	7.10	3.27	3.94
MATE		82.46	0.37	5.46	14.61	3.62	5.21
RBG1		83.14	1.04	2.98	8.73	-0.55	-0.03
RBG3		84.49	2.39	4.61	15.27	2.31	1.27
TurboParser		82.17	0.07	5.46	15.21	3.52	3.75
Languages							
German		82.42	0.32	—	—	—	—
English		87.05	4.95	1.44	0.44	11.29	14.99
Spanish		85.60	3.50	2.28	-0.44	2.39	10.99
Finnish		69.39	-12.71	-8.14	-2.26	-9.60	-22.68
Swedish		82.03	-0.06	-5.28	-6.63	4.80	4.08
POS/morph information							
long		81.14	-0.96	—	—	—	—
long+feat		81.79	-0.31	-0.14	-0.02	1.24	0.53
short		82.57	0.48	-1.70	0.80	-5.52	-3.99
Conversion							
HamleDT		81.89	-0.21	—	—	—	—
UniDepTB		82.77	0.67	-2.28	2.38	-6.00	-6.36
Interactions							
German	long+feat	83.85	1.75	—	—	—	—
German	short	81.69	-0.40	—	—	—	—
English	long+feat	85.47	3.37	-0.79	-0.13	-2.91	-3.13
English	short	88.26	6.16	1.25	-1.00	5.28	3.26
Spanish	long+feat	85.97	3.87	-0.66	-0.42	-2.50	-3.05
Spanish	short	85.33	3.23	0.94	2.25	3.46	2.55
Finnish	long+feat	71.89	-10.21	2.56	0.58	1.42	6.42
Finnish	short	68.16	-13.94	-2.38	-3.08	-2.00	-4.62
Swedish	long+feat	80.96	-1.14	-0.43	0.05	-2.22	-2.89
Swedish	short	82.84	0.74	1.55	0.53	4.35	3.24
MaltParser	long+feat	77.23	-4.87	—	—	—	—
MaltParser	short	75.97	-6.12	—	—	—	—
MaltOptimizer	long+feat	82.77	0.67	0.76	0.00	1.05	3.20
MaltOptimizer	short	77.77	-4.33	-1.73	-3.99	-2.64	-0.42
MATE	long+feat	84.21	2.11	1.52	0.71	2.61	4.39
MATE	short	81.37	-0.73	-0.07	-3.50	0.71	0.31
RBG1	long+feat	80.51	-1.59	0.29	-0.44	0.14	0.85
RBG1	short	84.34	2.25	5.26	15.89	4.67	2.52
RBG3	long+feat	82.30	0.20	0.46	0.92	0.28	0.70
RBG3	short	85.54	3.45	4.84	12.19	3.38	2.27
TurboParser	long+feat	83.84	1.74	1.15	0.17	2.54	5.28
TurboParser	short	80.95	-1.15	-0.48	-4.08	0.23	0.96
German	HamleDT	83.34	1.24	—	—	—	—
German	UniDepTB	80.17	-1.93	—	—	—	—
English	UniDepTB	91.35	9.25	6.56	8.76	6.57	14.96
Spanish	UniDepTB	83.29	1.19	-1.94	-1.86	-2.47	3.89
Finnish	UniDepTB	73.35	-8.75	8.76	3.66	11.45	26.26
Swedish	UniDepTB	83.30	1.20	5.98	9.23	6.29	12.04
size	English	—	—	0.53	-0.72	-1.07	-0.59
size	Spanish	—	—	0.57	-0.80	0.09	-0.19
size	Finnish	—	—	-1.14	1.72	0.52	-0.11
size	Swedish	—	—	1.47	1.95	0.37	1.22

Table 1: Factor analysis of all experiments

Appendix B: Factor Analysis (USD, RBG parser, subsets)

Property		Average LAS	Δ Avg LAS	OLS Loadings LAS	Coord	Subj	Obj
Global Average/Intercept		84.74	—	76.63	72.75	65.95	51.35
size		—	—	1.15	1.39	1.52	2.28
Parsers							
RBG1		83.97	-0.77	—	—	—	—
RBG3		85.51	0.77	1.82	6.56	3.64	1.60
Languages							
German		83.72	-1.01	—	—	—	—
English		89.10	4.36	0.14	-16.47	13.22	17.16
Spanish		87.08	2.34	2.47	-14.13	8.35	23.06
Finnish		69.99	-14.75	-2.87	-0.59	-1.91	-4.41
Swedish		84.56	-0.18	-2.61	-16.50	15.01	10.48
POS/morph information							
long		80.72	-4.02	—	—	—	—
long+feat		81.34	-3.40	0.79	-0.60	3.78	3.83
short		86.29	1.55	4.06	11.43	3.50	0.57
Conversion							
HamleDT		83.91	-0.83	—	—	—	—
UniDepTB		86.47	1.73	-2.69	-6.47	-0.76	3.31
Interactions							
German	long+feat	84.28	-0.46	—	—	—	—
German	short	83.68	-1.06	—	—	—	—
English	long+feat	85.41	0.67	-1.00	-2.42	-4.85	-2.20
English	short	90.31	5.57	1.32	4.23	3.77	1.28
Spanish	long+feat	86.07	1.33	-1.14	-0.69	-4.29	-3.47
Spanish	short	87.37	2.63	0.41	5.72	-0.21	0.92
Finnish	long+feat	69.03	-15.71	1.91	1.05	-1.74	-1.21
Finnish	short	74.68	-10.06	4.56	-0.02	6.32	2.09
Swedish	long+feat	81.93	-2.81	-0.89	0.89	-3.60	-3.21
Swedish	short	85.60	0.86	0.62	4.47	-0.85	1.90
RBG3	long	81.63	-3.11	—	—	—	—
RBG3	long+feat	82.31	-2.43	0.12	1.60	-0.05	-1.13
RBG3	short	86.98	2.24	-0.43	-3.47	-1.78	-0.64
English	UniDepTB	91.96	7.22	6.11	13.92	1.14	12.11
Spanish	UniDepTB	85.56	0.82	-0.91	5.03	-5.94	-5.79
Swedish	UniDepTB	86.17	1.43	5.47	12.83	3.41	9.08
size	English	—	—	0.74	1.97	0.09	-0.08
size	Spanish	—	—	0.51	1.79	0.63	-0.46
size	Finnish	—	—	-5.52	-1.12	-3.66	-8.47
size	Swedish	—	—	0.91	2.70	-0.63	0.72

size‡: $\log_2$ of $\frac{\text{sentences}}{1024}$

Table 2: Factor analysis: only "standard" Universal Dependencies, only RBG1/RBG3 but including 3 500 and 10 000 sentences reduced training sets