Do dependency parsing metrics correlate with human judgments?

Barbara Plank,[⋆] Héctor Martínez Alonso,[⋆] Željko Agić,[⋆] Danijela Merkler,[†] Anders Søgaard[⋆]
[⋆]Center for Language Technology, University of Copenhagen, Denmark
[†]Department of Linguistics, University of Zagreb, Croatia
bplank@cst.dk

Abstract

Using automatic measures such as labeled and unlabeled attachment scores is common practice in dependency parser evaluation. In this paper, we examine whether these measures correlate with human judgments of overall parse quality. We ask linguists with experience in dependency annotation to judge system outputs. We measure the correlation between their judgments and a range of parse evaluation metrics across five languages. The human-metric correlation is lower for dependency parsing than for other NLP tasks. Also, inter-annotator agreement is sometimes higher than the agreement between judgments and metrics, indicating that the standard metrics fail to capture certain aspects of parse quality, such as the relevance of root attachment or the relative importance of the different parts of speech.

1 Introduction

In dependency parser evaluation, the standard accuracy metrics—labeled and unlabeled attachment scores—are defined simply as averages over correct attachment decisions. Several authors have pointed out problems with these metrics; they are both sensitive to annotation guidelines (Schwartz et al., 2012; Tsarfaty et al., 2011), and they fail to say anything about how parsers fare on rare, but important linguistic constructions (Nivre et al., 2010). Both criticisms rely on the intuition that some parsing errors are more important than others, and that our metrics should somehow reflect that. There are sentences that are hard to annotate because they are ambiguous, or because they contain phenomena peripheral to linguistic theory, such as punctuation, clitics, or fragments. Manning (2011) discusses similar issues for part-of-speech tagging.

To measure the variable relevance of parsing errors, we present experiments with human judgment of parse output quality across five languages: Croatian, Danish, English, German, and Spanish. For the human judgments, we asked professional linguists with dependency annotation experience to judge which of two parsers produced the better parse. Our stance here is that, insofar experts are able to annotate dependency trees, they are also able to determine the quality of a predicted syntactic structure, which we can in turn use to evaluate parser evaluation metrics. Even though downstream evaluation is critical in assessing the usefulness of parses, it also presents non-trivial challenges in choosing the appropriate downstream tasks (Elming et al., 2013), we see human judgments as an important supplement to extrinsic evaluation.

To the best of our knowledge, no prior study has analyzed the correlation between dependency parsing metrics and human judgments. For a range of other NLP tasks, metrics have been evaluated by how well they correlate with human judgments. For instance, the standard automatic metrics for certain tasks—such as BLEU in machine translation, or ROUGE-N and NIST in summarization or natural language generation—were evaluated, reaching correlation coefficients well above .80 (Papineni et al., 2002; Lin, 2004; Belz and Reiter, 2006; Callison-Burch et al., 2007).

We find that correlations between evaluation metrics and human judgments are weaker for dependency parsing than other NLP tasks—our correlation coefficients are typically between .35 and .55—and that inter-annotator agreement is sometimes higher than human-metric agreement. Moreover, our analysis (§5) reveals that humans have a preference for attachment over labeling decisions, and that attachments closer to the root are more important. Our findings suggest that the currently employed metrics are not fully adequate.

Contributions We present i) a systematic comparison between a range of available dependency parsing metrics and their correlation with human judgments; and ii) a novel dataset[1] of 984 sentences (up to 200 sentences for each of the 5 languages) annotated with human judgments for the preferred automatically parsed dependency tree, enabling further research in this direction.

2 Metrics

We evaluate seven dependency parsing metrics, described in this section.

Given a labeled gold tree $G = \langle V, E_G, l_G(\cdot)\rangle$ and a labeled predicted tree $P = \langle V, E_P, l_P(\cdot)\rangle$, let $E \subset V \times V$ be the set of directed edges from dependents to heads, and let $l : V \times V \rightarrow L$ be the edge labeling function, with L the set of dependency labels.

The three most commonly used metrics are those from the CoNLL 2006–7 shared tasks (Buchholz and Marsi, 2006): unlabeled attachment score (UAS), label accuracy (LA), both introduced by Eisner (1996), and labeled attachment score (LAS), the pivotal dependency parsing metric introduced by Nivre et al. (2004).

$$\text{UAS} = \frac{|\{e \mid e \in E_G \cap E_P\}|}{|V|}$$

$$\text{LAS} = \frac{|\{e \mid l_G(e) = l_P(e), e \in E_G \cap E_P\}|}{|V|}$$

$$\text{LA} = \frac{|\{v \mid v \in V, l_G(v,\cdot) = l_P(v,\cdot)\}|}{|V|}$$

We include two further metrics—namely, labeled (LCP) and unlabeled (UCP) complete predications—to give account for the relevance of correct predicate prediction for parsing quality.

LCP is inspired by the *complete predicates* metric from the SemEval 2015 shared task on semantic parsing (Oepen et al., 2015).[2] LCP is triggered by a verb (i.e., set of nodes V_{verb}) and checks whether all its core arguments match, i.e., all outgoing dependency edges except for punctuation. Since LCP is a very strict metric, we also evaluate UCP, its unlabeled variant. Given a function $c_X(v)$ that retrieves the set of child nodes of a node v from a tree X, we first define UCP as follows, and then incorporate the label matching for LCP:

$$\text{UCP} = \frac{|\{v \mid V_{verb}, c_G(v) = c_P(v)\}|}{|V_{verb}|}$$

$$\text{LCP} = \frac{|\{v \mid V_{verb}, c_G(v) = c_P(v) \wedge l_G(v,\cdot) = l_P(v,\cdot)\}|}{|V_{verb}|}$$

For the final figure of seven different parsing metrics, on top of the previous five, in our experiments we also include the neutral edge direction metric (NED) (Schwartz et al., 2011), and tree edit distance (TED) (Tsarfaty et al., 2011; Tsarfaty et al., 2012).[3]

3 Experiment

In our analysis, we compare the metrics with human judgments. We examine how well the automatic metrics correlate with each other, as well as with human judgments, and whether inter-annotator agreement exceeds annotator-metric agreement.

LANG	TYPE	SENT	$\overline{SL}$	$\overline{TD}$	ANN	RAW	κ
da	CDT	200	22.7	8.1	2-3	.77	.53
de	UD	200	18.0	4.4	2	.67	.33
en	UD	200	23.4	5.4	4	.73	.45
es	UD	184	32.5	6.7	4	.60	.20
hr	PDT	200	28.5	7.8	2	.80	.59

Table 1: Data characteristics and agreement statistics. TD: tree depth; SL: sentence length.

Data In our experiments we use data from five languages: The English (en), German (de) and Spanish (es) treebanks from the Universal Dependencies (UD v1.0) project (Nivre et al., 2015), the Copenhagen Dependency Treebank (da) (Buch-Kromann, 2003), and the Croatian Dependency Treebank (hr) (Agić and Merkler, 2013). We keep the original POS tags for all datasets (17 tags in case of UD, 13 tags for Croatian, and 23 for Danish). Data characteristics are in Table 1.

For the parsing systems, we follow McDonald and Nivre (2007) and use the second order MST (McDonald et al., 2005), as well as Malt parser with pseudo-projectivization (Nivre and Nilsson, 2005) and default parameters. For each language, we train the parsers on the canonical training section. We randomly select 200 sentences from the test sections, where our two de-

[1] The dataset is publicly available at `https://bitbucket.org/lowlands/release`

[2] `http://alt.qcri.org/semeval2015/`

[3] `http://www.tsarfaty.com/unipar/`

Lang	Parser	LAS	UAS	LA	NED	TED	LCP	UCP
en	Malt	**79.17**	82.31	87.88	84.34	85.20	41.27	47.17
	MST	78.30	82.91	86.80	84.72	83.49	36.05	45.58
es	Malt	78.72	82.85	87.34	82.90	84.20	34.00	43.00
	MST	**79.51**	84.97	86.95	85.00	83.16	31.83	44.00
da	Malt	79.28	83.40	85.92	83.39	77.50	47.69	55.23
	MST	**82.75**	87.00	88.42	87.01	78.39	52.31	62.31
de	Malt	69.09	75.70	82.05	75.54	80.37	19.72	30.45
	MST	**72.07**	80.29	82.22	80.13	78.94	19.38	33.22
hr	Malt	63.21	72.34	76.66	71.94	71.64	23.18	31.03
	MST	**65.98**	76.20	79.01	75.89	72.82	24.71	34.29
Avg	Malt	73.84	79.32	83.97	76.62	**79.78**	**33.17**	43.18
	MST	**75.72**	**82.27**	**84.68**	**82.55**	79.36	32.86	**44.08**

Table 2: Parsing performance of Malt and MST.

pendency parsers do not agree on the correct analysis, after removing punctuation.[4] We do not control for predicted trees matching the gold standard.

Annotation task A total of 7 annotators were involved in the annotation task. All the annotators are either native or fluent speakers, and well-versed in dependency syntax analysis.

For each language, we present the selected 200 sentences with their two predicted dependency structures to 2–4 annotators and ask them to rank which of the two parses is better. They see graphical representations of the two dependency structures, visualized with the *What's Wrong With My NLP?* tool.[5] The annotators were not informed of what parser produced which tree, nor had they access to the gold standard. The dataset of 984 sentences is available at: `https://bitbucket.org/lowlands/release` (folder `CoNLL2015`).

4 Results

First, we perform a standard evaluation in order to see how the parsers fare, using our range of dependency evaluation measures. In addition, we compute correlations between metrics to assess their similarity. Finally, we correlate the measures with human judgements, and compare average annotator and human-system agreements.

Table 2 presents the parsing performances with respect to the set of metrics. We see that using LAS, Malt performs better on English, while MST performs better on the remaining four languages.

Table 3 presents Spearman's ρ between metrics across the 5 languages. Some metrics are strongly

[4]For Spanish, we had fewer analyses where the two parsers disagreed, i.e., 184.

[5]`https://code.google.com/p/whatswrong/`

ρ	UAS	LA	NED	TED	LCP	UCP
LAS	.755	.622	.743	.556	.236	.286
UAS	–	.436	.869	.512	.211	.342
LA	–	–	.436	.419	.206	.154
NED	–	–	–	.499	.216	.339
TED	–	–	–	–	.175	.219
LCP	–	–	–	–	–	.352

Table 3: Correlations between metrics.

ρ	en	es	da	de	hr	All
LAS	**.547**	**.478**	.297	.466	**.540**	**.457**
UAS	.541	.437	**.331**	.453	.397	.425
LA	.387*	.250*	.232	.310	.467	.324*
NED	.541	.469	.318	**.501**	.446	.448
TED	.372*	.404	.323	.331	.405*	.361*
LCP	.022*	.230*	.171	.120*	.120*	.126*
UCP	.249*	.195*	.223	.190*	.143*	.195*

Table 4: Correlations between human judgments and metrics (micro avg). * means significantly different from LAS ρ using Fisher's z-transform. Bold: highest correlation per language.

correlated, e.g., LAS and LA, and UAS and NED, but some exhibit very low correlation coefficients.

Next we study correlations with human judgments (Table 4). In order to aggregate over the annotations, we use an item-response model (Hovy et al., 2013). The correlations are relatively weak compared to similar findings for other NLP tasks. For instance, ROUGE-1 (Lin, 2004) correlates strongly with perceived summary quality, with a coefficient of 0.99. The same holds for BLEU and human judgments of machine translation quality (Papineni et al., 2002).

We find that, overall, LAS is the metric that correlates best with human judgments. It is closely followed by UAS, which does not differ significantly from LAS, albeit the correlations for UAS are slightly lower on average. NED is in turn highly correlated with UAS. The correlations for the predicate-based measures (LCP, UCP) are the lowest, as they are presumably too strict, and very different to LAS.

Motivated by the fact that people prefer the parse that gets the overall structure right (§5), we experimented with weighting edges proportionally to their log-distance to root. However, the signal was fairly weak; the correlations were only slightly higher for English and Danish: .552 and .338, respectively.

Finally, we compare the mean agreement be-

	ANN	LAS	UAS	LA	NED	TED	LCP	UCP
da	.768	.838	**.848**	.808	.828	.828	.745	.765
de	.670	**.710**	.690	.635	.710	.630	.575	.565
en	**.728**	.715	.705	.660	.700	.658	.525	.600
es	.601	**.663**	.644	.603	.652	.635	.581	.554
hr	**.800**	.755	.700	.735	.730	.705	.570	.580

Table 5: Average mean agreement between annotators, and between annotators and metrics.

tween humans with the mean agreement between humans and standard metrics, cf. Table 5. For two languages (English and Croatian), humans agree more with each other than with the standard metrics, suggesting that metrics are not fully adequate. The mean agreement between humans is .728 for English, with slightly lower scores for the metrics (LAS: .715, UAS: .705, NED: .660). The difference between mean agreement of annotators and human-metric was higher for Croatian: .80 vs .755. For Danish, German and Spanish, however, average agreement between metrics and human judgments is higher than our inter-annotator agreement.

5 Analysis

In sum, our experiments show that metrics correlate relatively weakly with human judgments, suggesting that some errors are more important to humans than others, and that the relevance of these errors are not captured by the metrics.

To better understand this, we first consider the POS-wise correlations between human judgments and LAS, cf. Table 6. In English, for example, the correlation between judgments and LAS is significantly stronger for content words[6] ($\rho_c = 0.522$) than for function words ($\rho_f = 0.175$). This also holds for the other UD languages, namely German ($\rho_c = 0.423$ vs $\rho_f = 0.263$) and Spanish ($\rho_c = 0.403$ vs $\rho_f = 0.228$). This is not the case for the non-UD languages, Croatian and Danish, where the difference between content-POS and function-POS correlations is not significantly different. In Danish, function words head nouns, and are thus more important than in UD, where content-content word relations are annotated, and function words are leaves in the dependency tree. This difference in dependency formalism is shown by the higher correlation for ρ_f for Danish.

The greater correlation for content words for English, German and Spanish suggests that errors

[6]Tagged as ADJ, NOUN, PROPN, VERB.

ρ	content	function
en	.522	.175
de	.423	.263
es	.403	.228
da	.148	.173
hr	.340	.306

Table 6: Correlations between human judgements and POS-wise LAS (content ρ_c vs function ρ_f poswise LAS correlations).

in attaching or labeling content words mean more to human judges than errors in attaching or labeling function words. We also observe that longer sentences do not compromise annotation quality, with a ρ between -0.07 and 0.08 across languages regarding sentence length and agreement.

For the languages for which we had 4 annotators, we analyzed the subset of trees where humans and system (by LAS) disagreed, but where there was majority vote for one tree. We obtained 35 dependency instances for English and 27 for Spanish (cf. Table 7). Two of the authors determined whether humans preferred labeling over attachment, or otherwise.

	attachment	labeling	items
en	86%	14%	35
es	67%	33%	27

Table 7: Preference of attachment or labeling for items where humans and system disagreed and human agreement ≥ 0.75.

Table 7 shows that there is a prevalent preference for attachment over labeling for both languages. For Spanish, there is proportionally higher label preference. Out of the attachment preferences, 36% and 28% were related to root/main predicate attachments, for English and Spanish respectively. The relevance of the root-attachment preference indicates that attachment is more important than labeling for our annotators.

Figure 5 provides three examples from the data where human and system disagree. Parse i) involves a coordination as well as a (local) adverbial, where humans voted for correct coordination (red) and thus unanimously preferred attachment over labeling. Yet, LAS was higher for the analysis in blue because "certainly" is attached to "Europeans" in the gold standard. Parse ii) is another

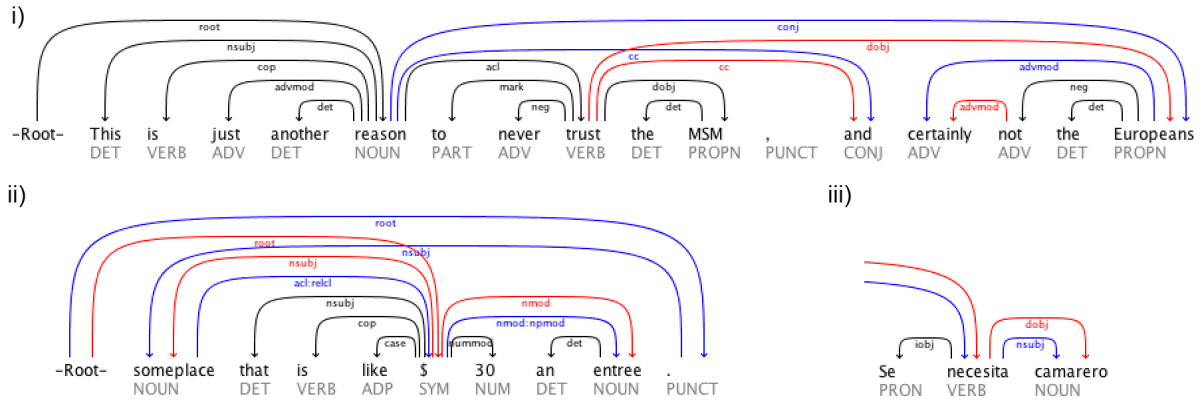

Figure 1: Examples where human and system (LAS) disagree. Human choice: i) red; ii) red; iii) blue.

example where humans preferred attachment (in this case root attachment), while iii) shows a Spanish example ("waiter is needed") where the subject label (*nsubj*) of "camarero" ("waiter") was the decisive trait.

6 Related Work

Parsing metrics are sensitive to the choice of annotation scheme (Schwartz et al., 2012; Tsarfaty et al., 2011) and fail to capture how parsers fare on important linguistic constructions (Nivre et al., 2010). In other NLP tasks, several studies have examined how metrics correlate with human judgments, including machine translation, summarization and natural language generation (Papineni et al., 2002; Lin, 2004; Belz and Reiter, 2006; Callison-Burch et al., 2007). Our study is the first to assess the correlation of human judgments and dependency parsing metrics. While previous studies reached correlation coefficients over 0.80, this is not the case for dependency parsing, where we observe much lower coefficients.

7 Conclusions

We have shown that out of seven metrics, LAS correlates best with human jugdments. Nevertheless, our study shows that there is an amount of human preference that is not captured with LAS. Our analysis on human versus system disagreement indicates that attachment is more important than labeling, and that humans prefer a parse that gets the overall structure right. For some languages, inter-annotator agreement is higher than annotator-metric (LAS) agreement, and content-POS is more important than function-POS, indicating there is an amount of human preference that is not captured with our current metrics. These observations raise the important question on how to incorporate our observations into parsing metrics that provide a better fit to human judgments. We do not propose a better metric here, but simply show that while LAS seems to be the most adequate metric, there is still a need for better metrics to complement downstream evaluation.

We outline a number of extensions for future research. Among those, we would aim at augmenting the annotations by obtaining more detailed judgments from human annotators. The current evaluation would ideally encompass more (diverse) domains and languages, as well as the many diverse annotation schemes implemented in various publicly available dependency treebanks that were not included in our experiment.

Acknowledgments

We thank Muntsa Padró and Miguel Ballesteros for their help and the three anonymous reviewers for their valuable feedback.

References

Željko Agić and Danijela Merkler. 2013. Three Syntactic Formalisms for Data-Driven Dependency Parsing of Croatian. In *Text, Speech, and Dialogue*. Springer.

Anja Belz and Ehud Reiter. 2006. Comparing Automatic and Human Evaluation of NLG Systems. In *EACL*.

Matthias Buch-Kromann. 2003. The Danish Dependency Treebank and the DTAG Treebank Tool. In *TLT*.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. In *CoNLL*.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*.

Jason Eisner. 1996. Three new probabilistic models for dependency parsing. In *COLING*.

Jakob Elming, Anders Johannsen, Sigrid Klerke, Emanuele Lapponi, Héctor Martínez Alonso, and Anders Søgaard. 2013. Down-stream effects of tree-to-dependency conversions. In *NAACL*.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *NAACL*.

Chin-Yew Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*.

Christopher D Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Computational Linguistics and Intelligent Text Processing*. Springer.

Ryan McDonald and Joakim Nivre. 2007. Characterizing the errors of data-driven dependency parsers. In *EMNLP-CoNLL*.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *ACL*.

Joakim Nivre and Jens Nilsson. 2005. Pseudo-projective dependency parsing. In *ACL*.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2004. Memory-based dependency parsing. In *CoNLL*.

Joakim Nivre, Laura Rimell, Ryan McDonald, and Carlos Gomez-Rodriguez. 2010. Evaluation of dependency parsers on unbounded dependencies. In *COLING*.

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.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinkova, Dan Flickinger, Jan Hajic, and Zdenka Uresova. 2015. Semeval 2015 task 18: Broad-coverage semantic dependency parsing. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

Kishore Papineni, Salim Roukus, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*, Philadelphia, Pennsylvania.

Roy Schwartz, Omri Abend, Roi Reichart, and Ari Rappoport. 2011. Neutralizing linguistically problematic annotations in unsupervised dependency parsing evaluation. In *ACL*.

Roy Schwartz, Omri Abend, and Ari Rappoport. 2012. Learnability-based syntactic annotation design. In *COLING*.

Reut Tsarfaty, Joakim Nivre, and Evelina Andersson. 2011. Evaluating dependency parsing: robust and heuristics-free cross-annotation evaluation. In *EMNLP*.

Reut Tsarfaty, Joakim Nivre, and Evelina Andersson. 2012. Cross-framework evaluation for statistical parsing. In *EACL*.