Do dependency parsing metrics correlate with human judgments?

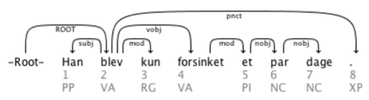
Barbara Plank, Héctor Martínez Alonso, Željko Agić, Danijela Merkler,♦ Anders Søgaard

Center for Language Technology, University of Copenhagen, Denmark
♦Department of Linguistics, University of Zagreb, Croatia

bplank@cst.dk

- A systematic comparison between 7 dependency parsing evaluation metrics and human judgments of overall parse quality.
- A novel dataset of 984 sentences annotated with human judgments for five languages.
  - Human-metric correlation is lower for dependency parsing than for other NLP tasks.
  - Inter-annotator agreement is sometimes higher than agreement between judgments and metrics.
  - Humans have a preference for attachment over labeling, and attachment closer to the root is more important.

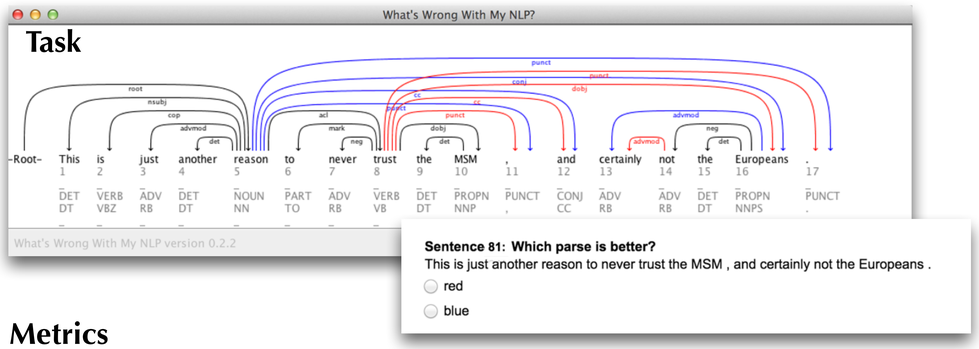
Datasets



	Annotator	*LAS*	*UAS*	*LA*	*NED*	*TED*	*LCP*	*UCP*
da	0.768	0.838	**0.848**	0.808	0.828	0.828	0.745	0.765
de	0.670	**0.710**	0.690	0.635	0.710	0.630	0.575	0.565
en	**0.728**	0.715	0.705	0.660	0.700	0.658	0.525	0.600
es	0.601	**0.663**	0.644	0.603	0.652	0.635	0.581	0.554
hr	**0.800**	0.755	0.700	0.730	0.730	0.705	0.570	0.580

Average mean agreement between annotators, and between annotators and metrics.

Task



Metrics

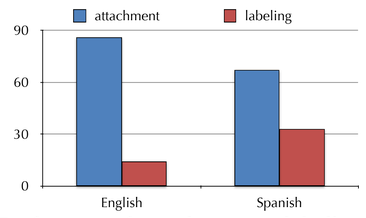
- Unlabeled attachment score (**UAS**)
- Labeled attachment score (**LAS**)
- Label accuracy (**LA**)
- Unlabeled complete predicates (**UCP**)
- Labeled complete predicates (**LCP**)
- Neutral Edge Direction (**NED**) (Schwartz et al., 2011)
- Tree Edit Distance (**TED**) (Tsarfaty et al. 2011; 2012)

Results

ρ	*English*	*Spanish*	*Danish*	*German*	*Croatian*	*All*
LAS	**0.547**	**0.478**	0.297	0.466	**0.540**	**0.457**
UAS	0.541	0.437	**0.331**	0.453	0.397	0.425
LA	0.387*	0.250*	0.232	0.310	0.467	0.324*
NED	0.541	0.469	0.318	**0.501**	0.446	0.448
TED	0.372*	0.404	0.323	0.331	0.405*	0.361*
LCP	0.022*	0.230*	0.171	0.120*	0.120*	0.126*
UCP	0.249*	0.195*	0.223	0.190*	0.143*	0.195*

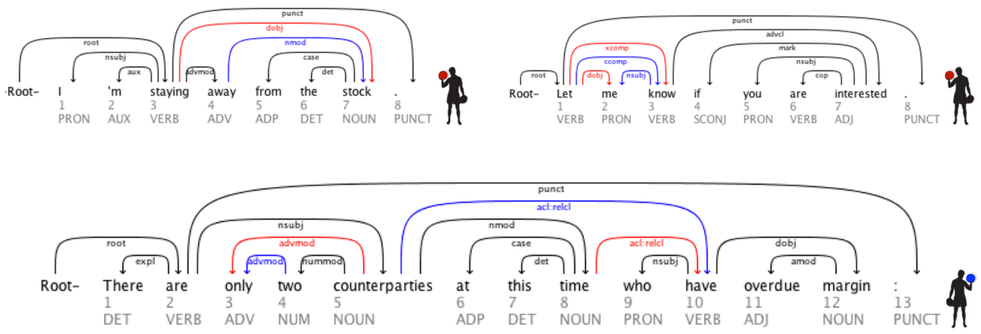
Correlations between human judgments and metrics. Bold: highest correlation per language. *means significantly different from LAS.

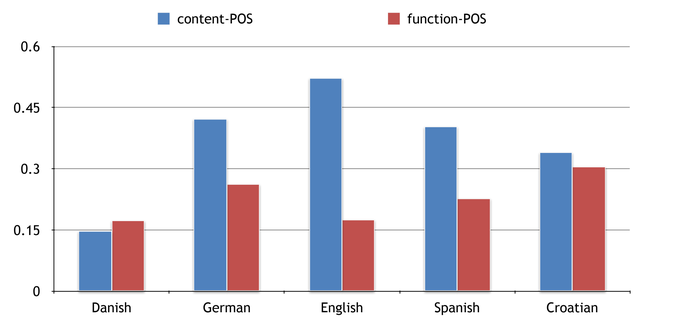
Analysis



Preference of attachment or labeling for items where human and system disagree and human agreement ≥ .75.

Examples





Correlations between human jugdments and POS-wise LAS.

CDT — UD — PDT