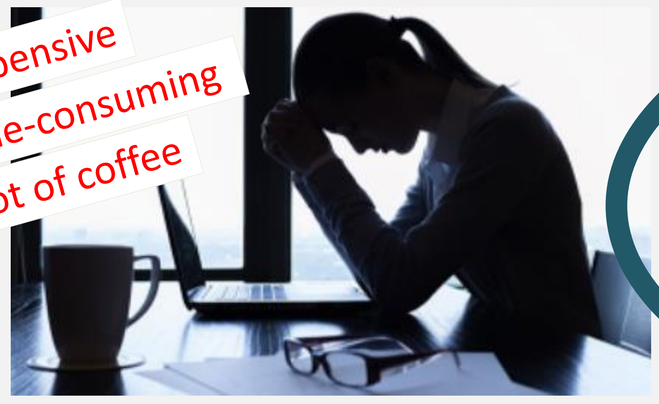
Quality and Efficiency of Manual Annotation: Pre-annotation Bias

Marie Mikulová, Milan Straka, Jan Štěpánek, Barbora Štěpánková, Jan Hajič
{mikulova,straka,stepanek,stepankova,hajic@ufal.mff.cuni.cz}
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

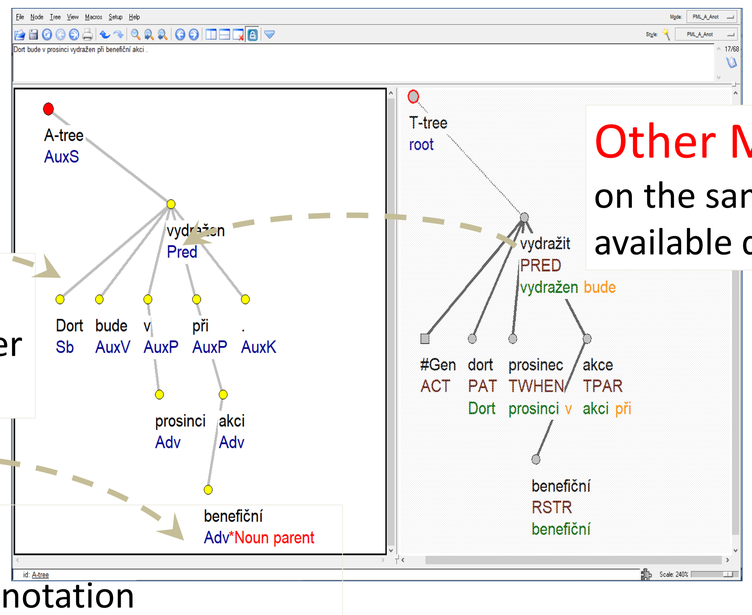
Dependency Syntax Annotation

- within multi-layered treebank for Czech: Prague Dependency Treebank – Consolidated 2.0
- structure and syntactic function label (25 types): Predicate, Subject, Object, Attribute, Adverbial etc.
- plus additional features: ellipsis, parenthesis, member of coordination (25x23=200 types of labels)
- 2,000,000 tokens to be annotated



Does the quality of manual annotation remain acceptably high if pre-annotation is used?

Annotation Support Tools



Pre-annotation
- high-accuracy parser
- UDPipe2

Checking Rules
- available during annotation
- e.g. node depends on a noun gets *Atr*

Other Manual Annotation
on the same data
available during annotation

Experiment Design

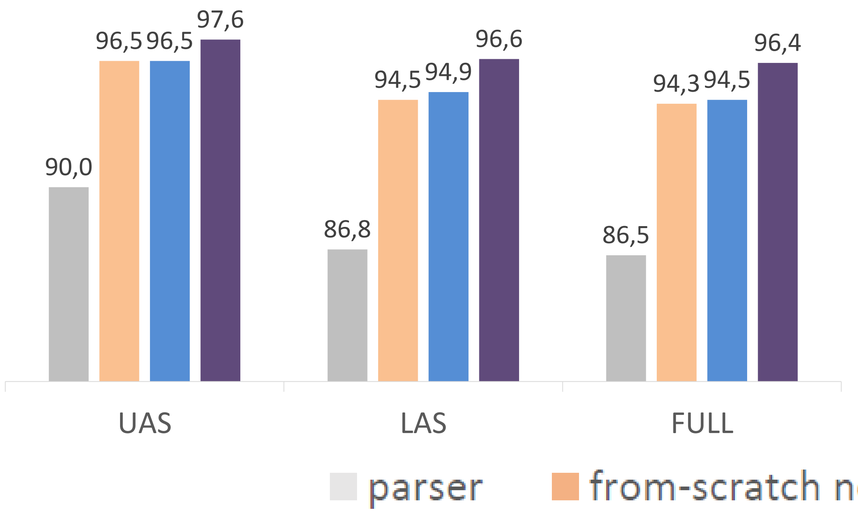
- data effect
- annotator effect
- learning effect

→ eliminated!

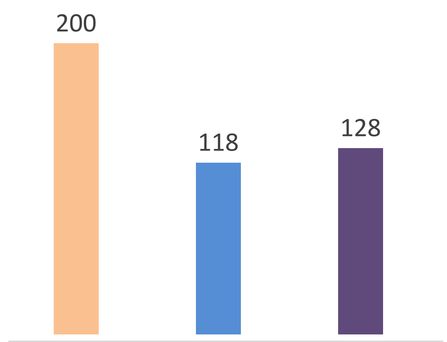
		Task							
		No Support		Other Annotation		Checking Rules		Other Annotation + Rules	
Mode	Pre-parsed	A1	B1	A1	B1	A1	B1	A1	B1
		A2	B2	A2	B2	A2	B2	A2	B2
	From-scratch	B1	A1	B1	A1	B1	A1	B1	A1
		B2	A2	B2	A2	B2	A2	B2	A2
		Data1 1,250 tokens	Data2 1,250 tokens	Data3 1,250 tokens	Data4 1,250 tokens	Data5 1,250 tokens	Data6 1,250 tokens	Data7 1,250 tokens	Data8 1,250 tokens
		Dataset							

Results

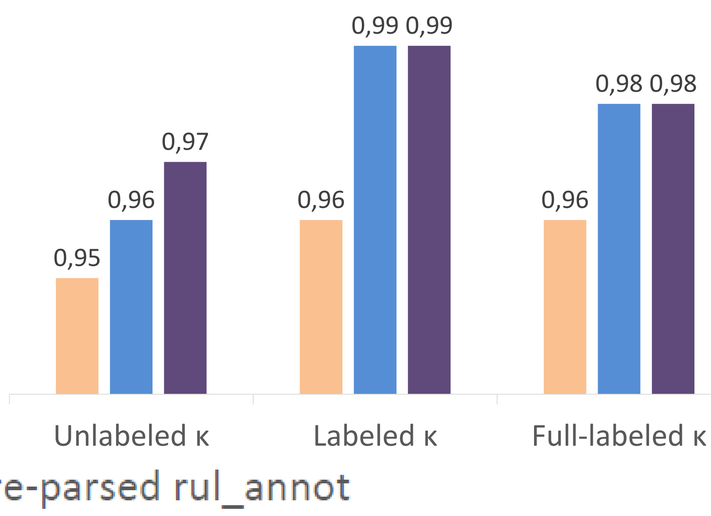
Accuracy



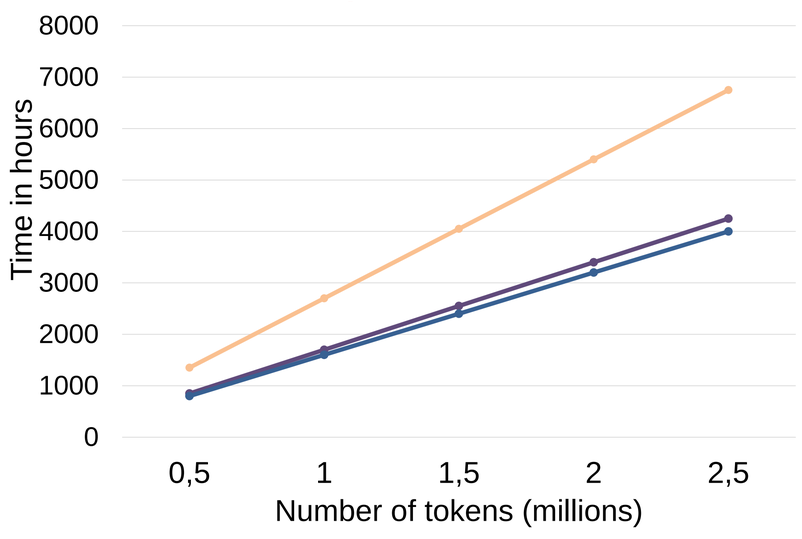
Time (in mins)



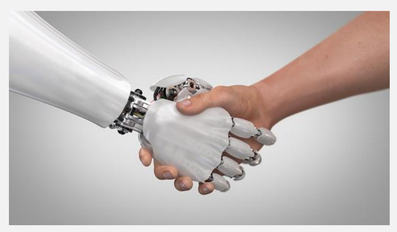
Inter-annotator agreement



Extrapolated hours



Conclusions



Automatic pre-annotation
- efficient tool for manual syntactic annotation
- increase in speed and consistency of annotation
- no reduction in quality

Use of other support tools
- together they significantly increase the quality of annotation
- no increase in annotation time needed (much)

Data from experiment available: http://hdl.handle.net/11234/1-4647

Supported by the LINDAT/CLARIAH-CZ Research Infrastructure, Ministry of Education CZ, LM2018101 & by project LUSyD, Grant Agency of the Czech Republic, GX20-16819X.