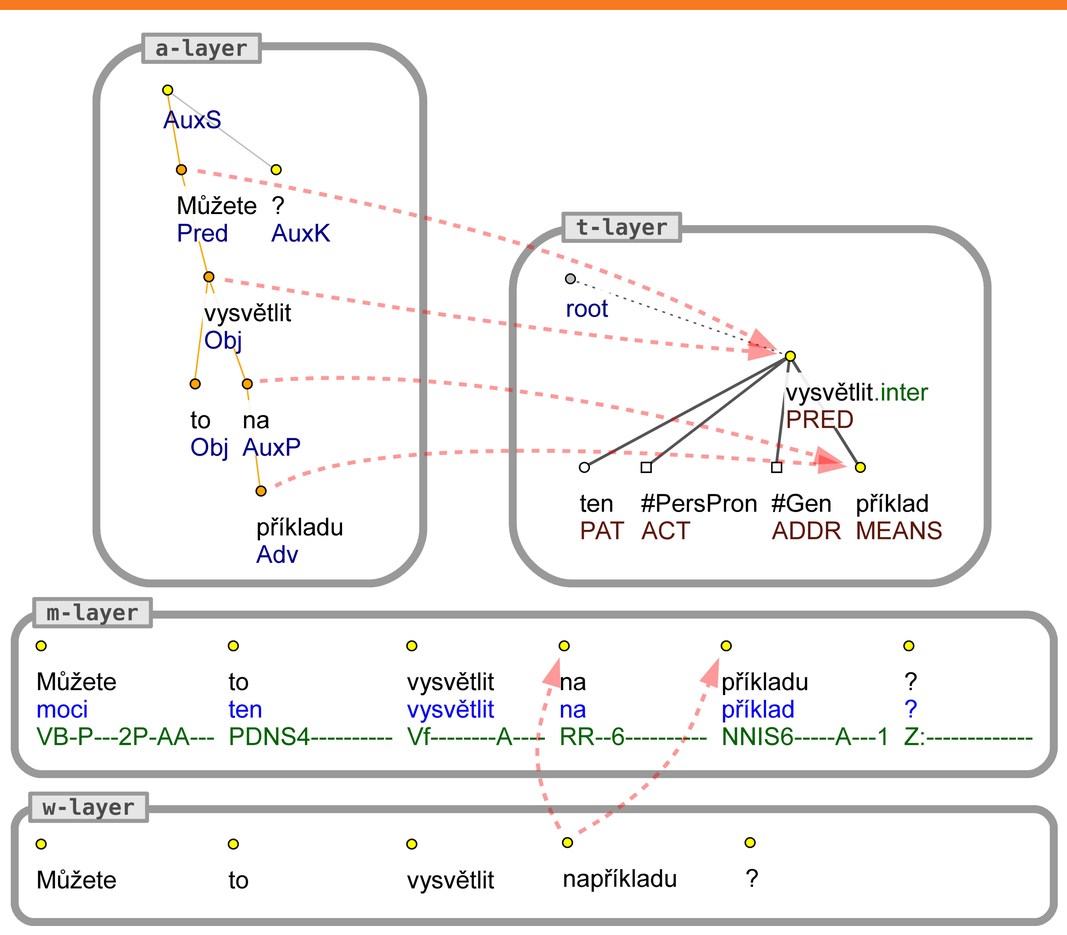
What can linguists learn from some simple statistics on annotated treebanks

Jiří Mírovský and Eva Hajičová
Charles University in Prague, Institute of Formal and Applied Linguistics

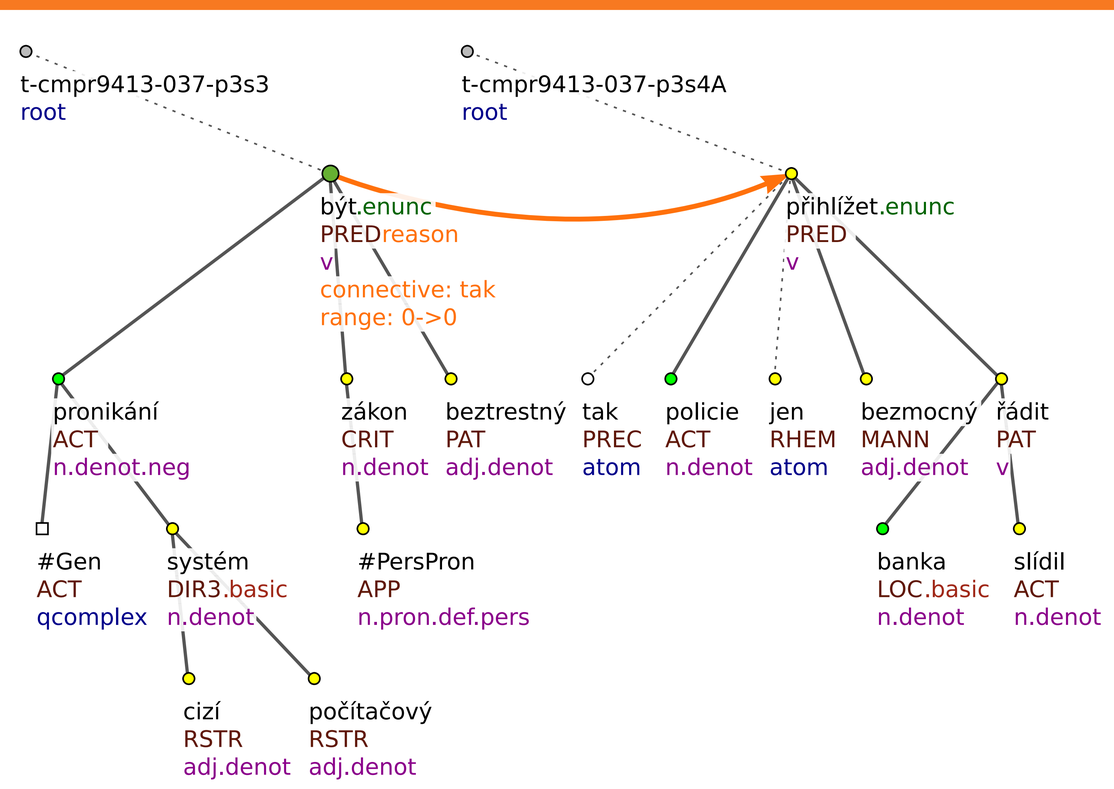
> "Une théorie inexacte amène une rectification, tandis que l'absence de théorie n'amène rien."
> *Maurice Grammont, as quoted by R. Jakobson*

Prague Dependency Treebank



Můžete to vysvětlit na příkladu?
[Can-you it explain on an-example?]
[Can you explain it on an example?]

Discourse Relations in Prague Dependency Treebank



*Pronikání do cizích počítačových systémů je podle našich zákonů beztrestné. Policie **tak** jen bezmocně přihlíží, když v bankách řádí slídilové.*

*[Infiltration into other computer systems is according to our laws not a criminal act. **Thus** the police only helplessly watches, as snoopers rage in banks.]*

Inter-Annotator Agreement

Inter-annotator agreement on various levels of annotation in PDT

Layer of annotation	IAA
Morphology (Czech, 5 thousand tags)	97%
Morphology (German, 54 tags)	98.57%
Surface syntax (German, unlabelled structural annotation)	92.43%
Surface syntax (German, labelled structural annotation (labelled nodes with 25 phrase types and labelled edges with 45 grammatical functions))	88.53%
Deep syntax – tectogrammatics (Czech, unlabelled structural annotation)	91%
Deep syntax – tectogrammatics (Czech, agreement on assigning the correct type to the dependency relation (67 functors))	84%
Topic-focus articulation (Czech, agreement on assigning contextual boundness to tectogrammatical nodes)	82%
Discourse relations (Czech, agreement on recognizing a presence of a discourse relation)	83%
Discourse relations (Czech, agreement on assigning one of 23 types)	77%
Textual coreference (Czech, agreement on recognizing a presence)	72%
Textual coreference (Czech, agreement on assigning one of two types)	90%
Bridging anaphora (Czech, agreement on recognizing a presence)	46%
Bridging anaphora (Czech, agreement on assigning one of 9 types)	92%
Genres of documents (Czech, agreement on assigning one of 20 genres)	77%

decrease in agreement:

morphology → discourse
coreference → bridging
types : presence ?

Frequencies

Discourse relations occurring intra-sententially and inter-sententially

Type of relation	Intra-sentential (%)	Inter-sentential (%)
Purpose	100	0
Condition	99	1
Disjunctive alternative	95	5
Pragmatic condition	93	7
Specification	82	18
Conjunction	81	19
Conjunctive alternative	79	21
Synchrony	77	23
Correction	73	27
Concession	70	30
Asynchrony	70	30
Reason – result	61	39
Confrontation	53	47
Gradation	52	48
Pragmatic opposition	46	54
Opposition	43	57
Explication	43	57
Equivalence	40	60
Restrictive opposition	37	63
Pragmatic reason – result	31	69
Exemplification	19	81
Generalisation	9	91

Discourse relations expressed by connectives and by AltLexes

	All	Intra-sentential	Inter-sentential
AltLex:	**726 (3.9%)**	**272 (2.1%)**	**454 (7.7%)**
Connective:	17,983 (96.1%)	12,523 (97.9%)	5,460 (92.3%)
Total:	18,709 (100%)	12,795 (100%)	5,914 (100%)

Pairs of non-sentential contextually non-bound participants

Expected systemic ordering:

Actor – Temporal (when – since when – till when – how long) – Location (where) – Manner – Extent – Measure – Means – Addressee - From where – Patient – To where – Effect – Condition – Aim - Cause

Parts of distributions of genres in five parts of PDT annotated by five annotators

A1		A2		A3		A4		A5	
428	news	147	news	118	description	179	news	157	news
124	description	50	comment	59	comment	43	sport	40	sport
113	sport	36	sport	41	sport	26	description	29	caption
68	essay	22	description	35	essay	22	review	28	description
60	other	20	caption	28	news	17	comment	23	comment

Final Remarks

Simple frequencies help confirm or disprove theories, reveal interesting directions for further research.

IAA measurements show difficulty of various tasks and improve the quality of annotations.

The Prague Dependency Treebank 3.0 was published in December 2013 under the Creative Commons License and it is available to download from the LINDAT-Clarin repository.

The presented work was supported by the Grant Agency of the Czech Republic (project n. P406/12/0658), and by the Ministry of Education, Youth and Sports of the Czech Republic (project n. LM2010013 – LINDAT/CLARIN).