

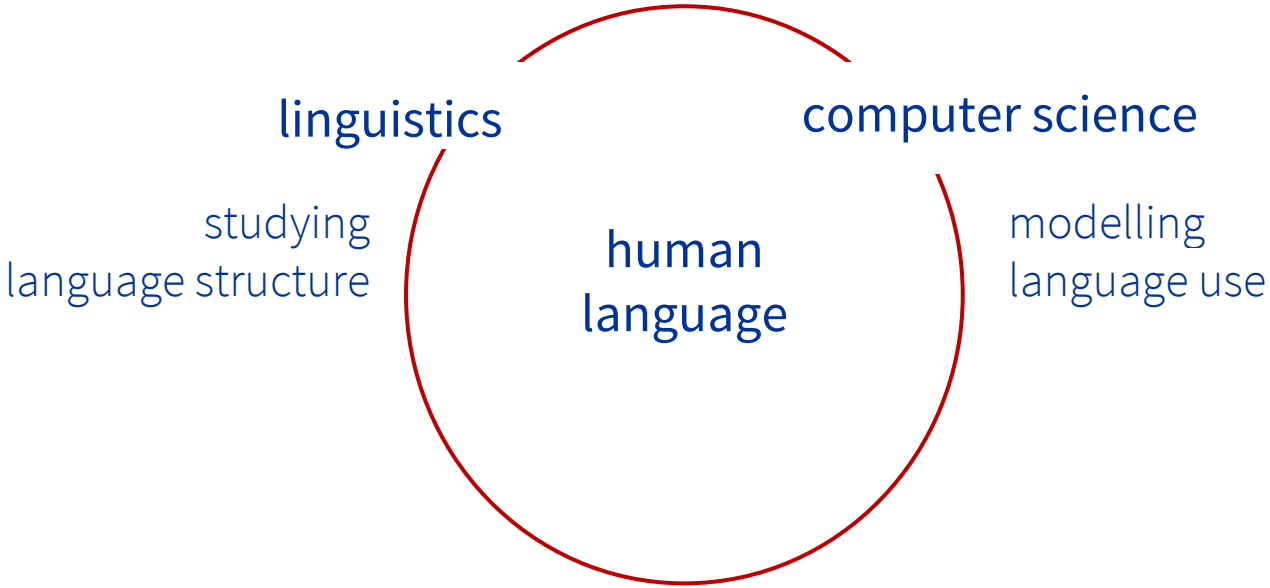

CHARLES UNIVERSITY
Faculty of mathematics
and physics

Creating and Exploiting Annotated Corpora

Barbora Vidová Hladká

June 3, 2020

Computational linguistics



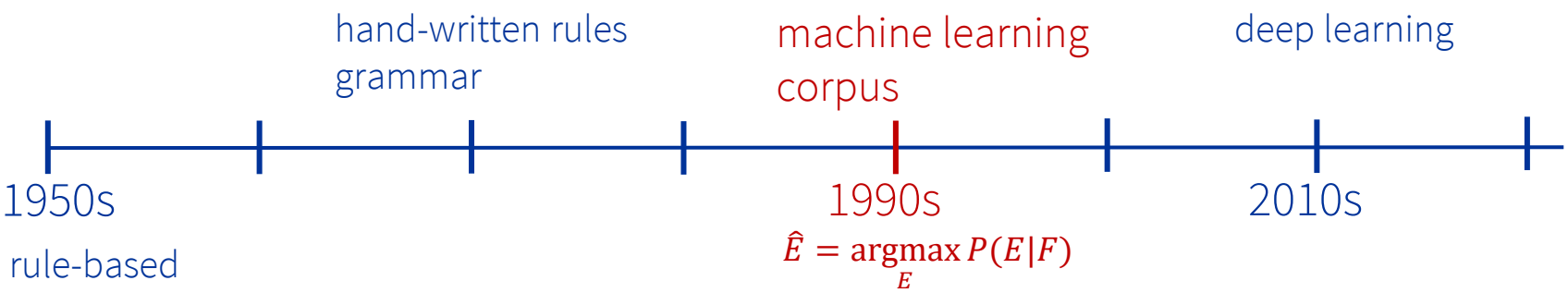
Computational linguistics

rule-based
machine translation
(Wiever, 1949)

statistical
machine translation
(Brown et al., 1990)

neural
machine translation
(Goodfellow et al., 2016)

Computational linguistics



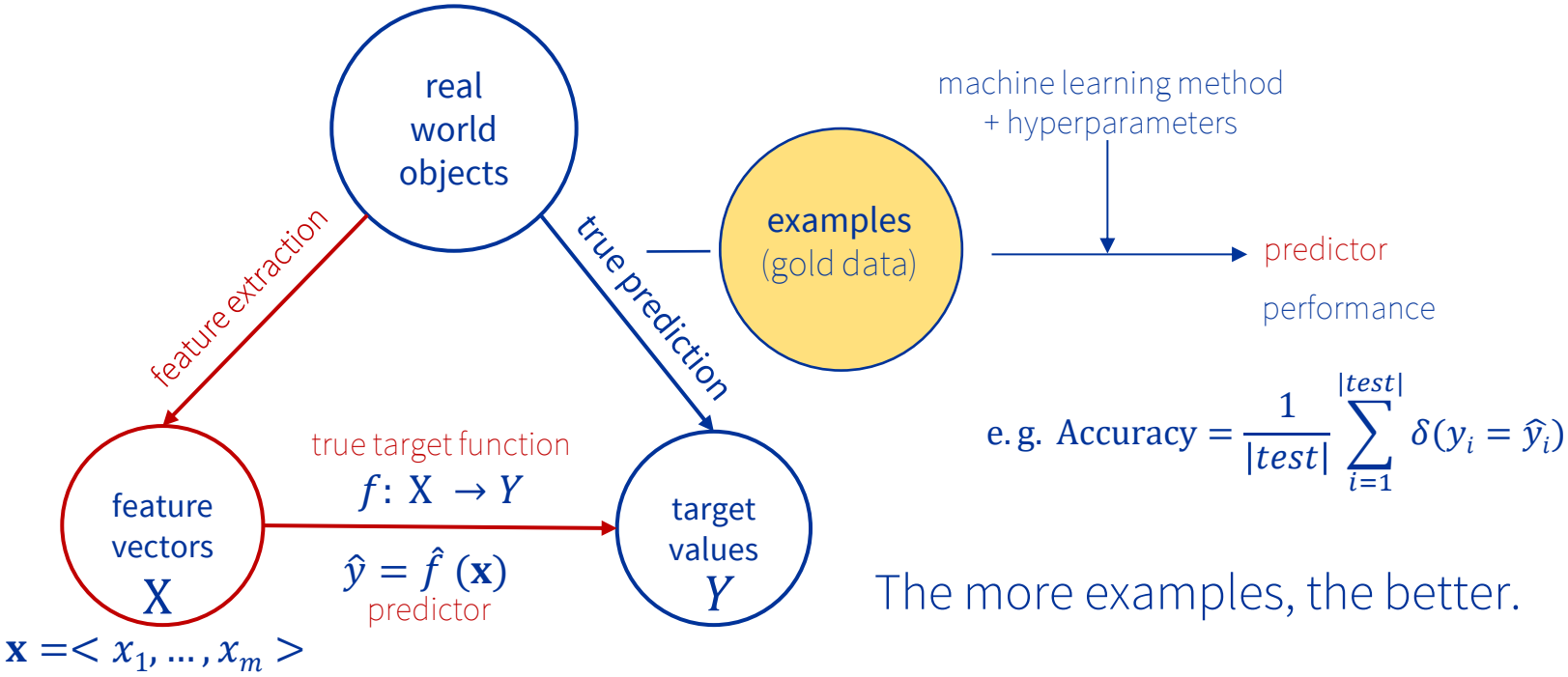
When we had our first results we submitted them to Coling 1988. Here is a part of the rejection review we received:

> The validity of a statistical (information theoretic) approach to MT has indeed been recognized, as the authors mention, by Weaver as early as 1949. And was universally recognized as mistaken by 1950 (cf. Hutchins, MT – Past, Present, Future, Ellis Horwood, 1986, p. 30ff and references therein). The crude force of computers is not science. The paper is simply beyond the scope of COLING.
>
> Anonymous Coling review, 1 March 1988

Frederick Jelinek, ACL Lifetime Achievement Award, 2009

Supervised machine learning

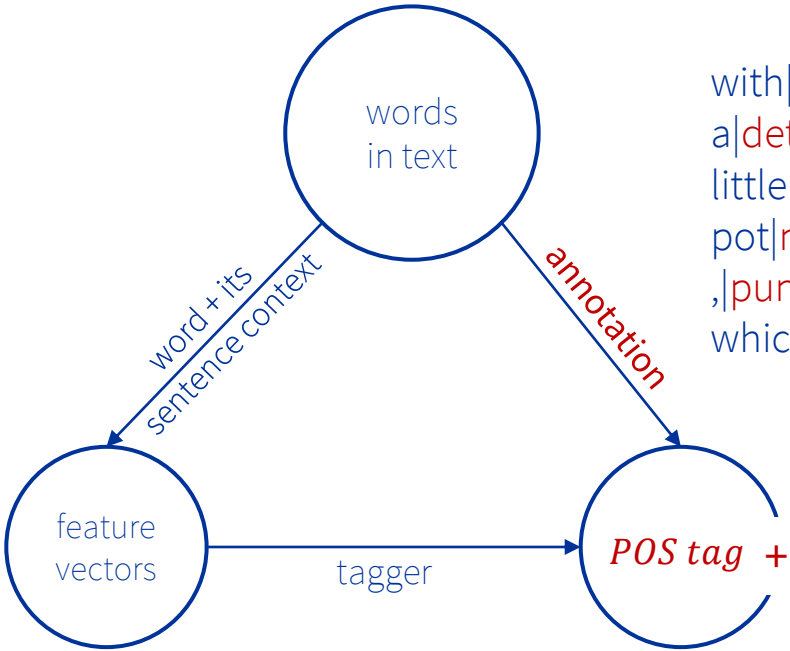
Learning from examples



The more examples, the better.

Supervised machine learning

Task: Label words in a text with their Part-Of-Speech classes (tagging)



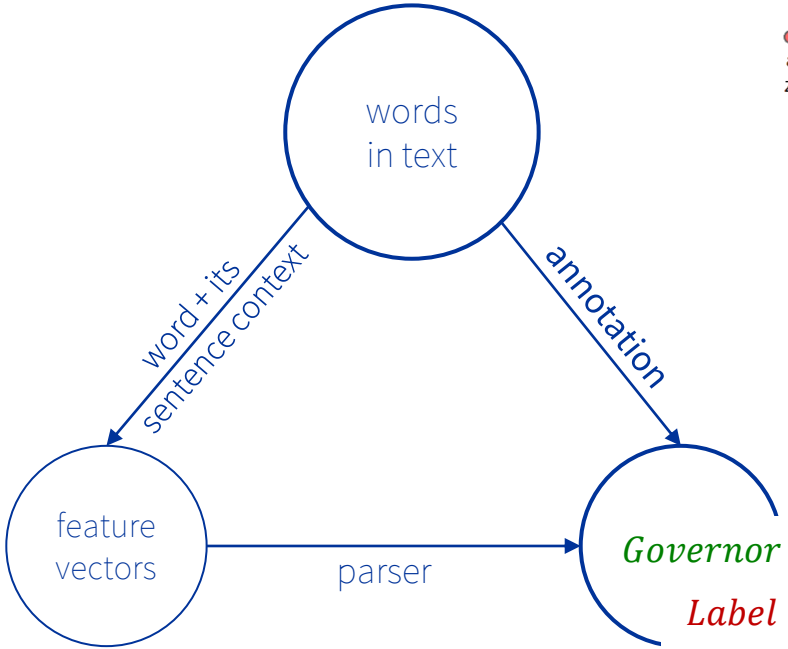
with|preposition
a|determiner
little|adjective
pot|noun
,|punctuation
which|pronoun

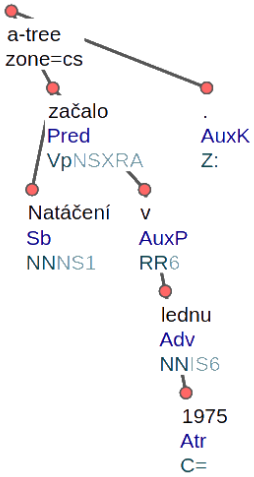
+ guidelines = annotation scheme

$\mathbf{x} = < preceding\ word, preceding\ POS\ tag, word\ ending, \ldots >$

Supervised machine learning

Task: Analyse the structure of a sentence (parsing)



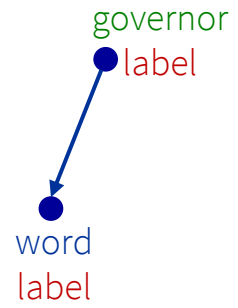


$Governor =$ {words in the sentence}

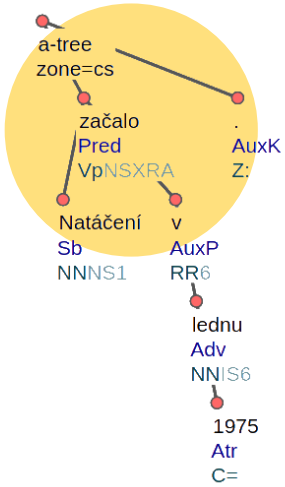
$Label =$ {Subject, Predicate, …}

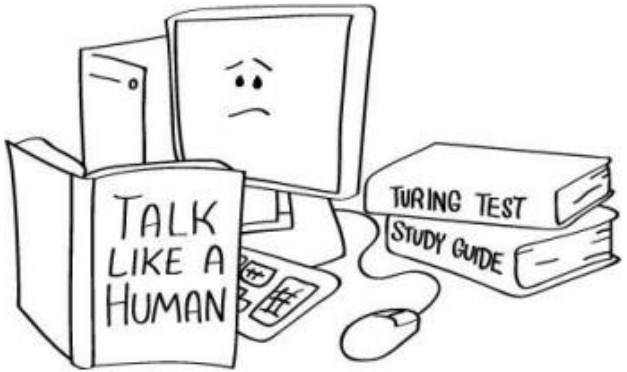
$\mathbf{x} = < preceding\ word, preceding\ POS\ tag, following\ POS\ tag, \ldots >$

syntactic dependency



Annotated corpus

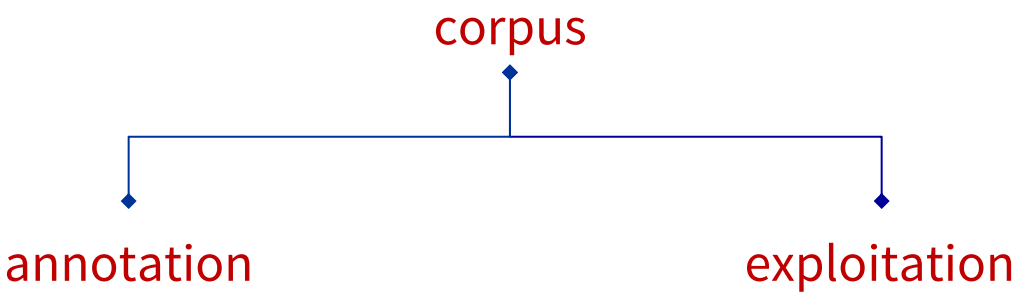




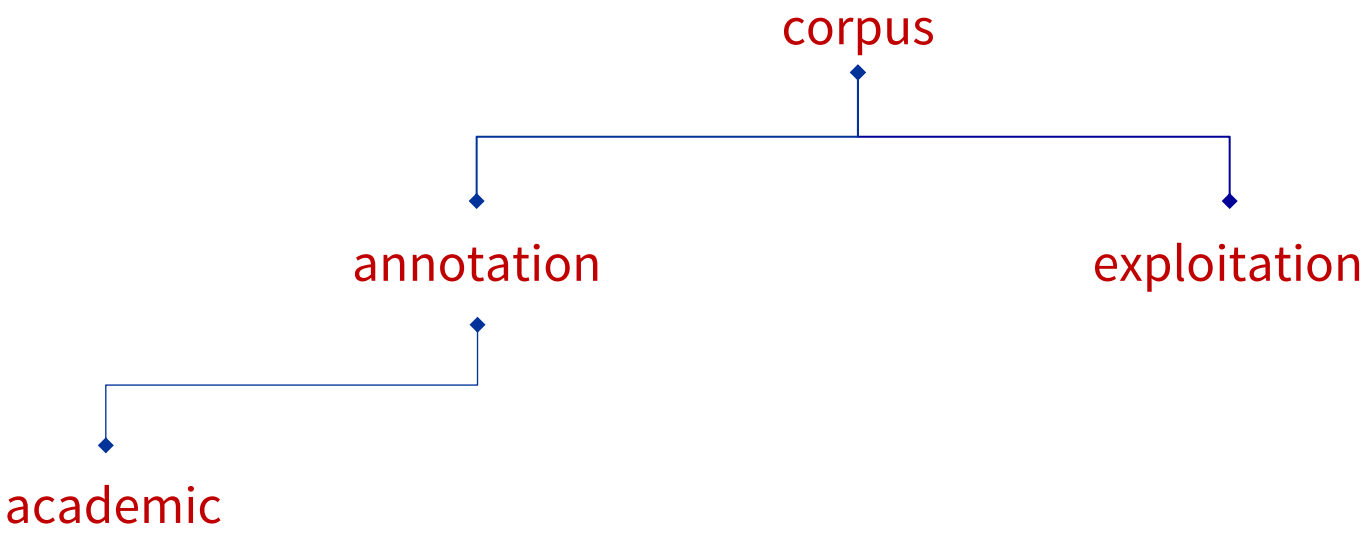
Exceptional position of Czech

- Brown corpus - English, 1M words, morph (Francis and Kucera, 1964)
- **Czech Academic Corpus - Czech, 540K words, morph+synt** (Těšitelová, 1985)
- Talbanken Corpus - Swedish, 350K words, morph+synt (Einarsson, 1976)
- …
- Penn Treebank - English, 2M words, morph+synt (Taylor et al., 2003)
- **Prague Dependency Treebank 3.5 - Czech, 2M words, morph+synt+sem** (Hajič et al., 2018)
- …

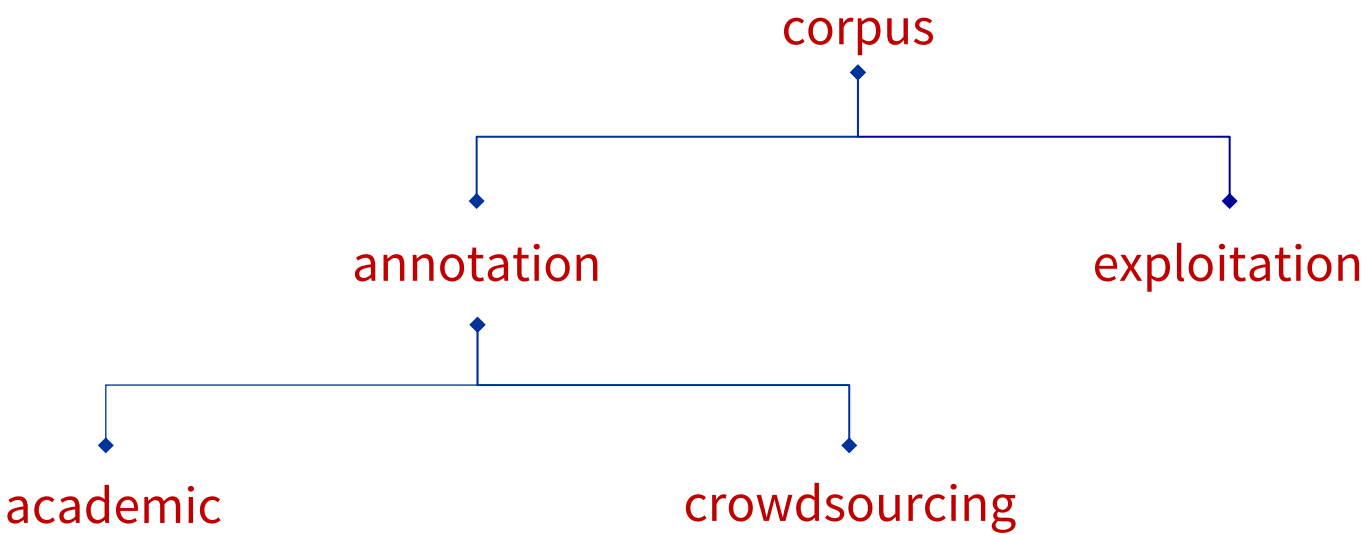
My research areas



My research areas

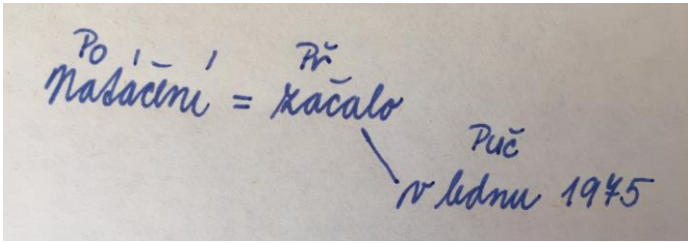


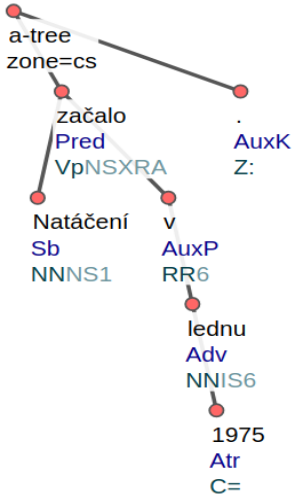
My research areas



Topic #1

Can we engage a crowd of Czech students into a faster and cheaper academic annotation?

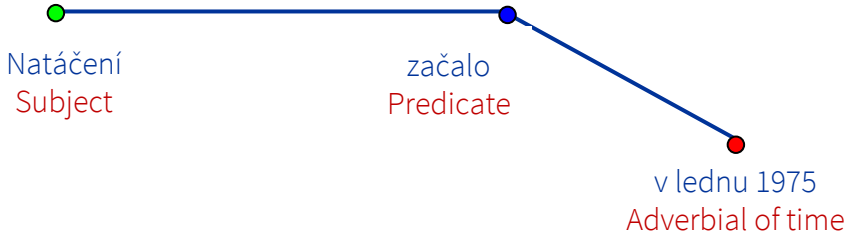


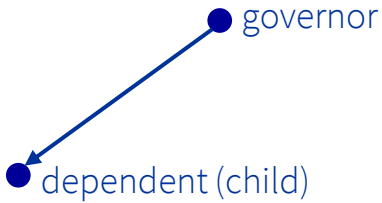


Sentence diagram over sentence (Hana, Hladká, 2014)

Example

Natáčení začalo v lednu 1975. Filming began in January 1975.





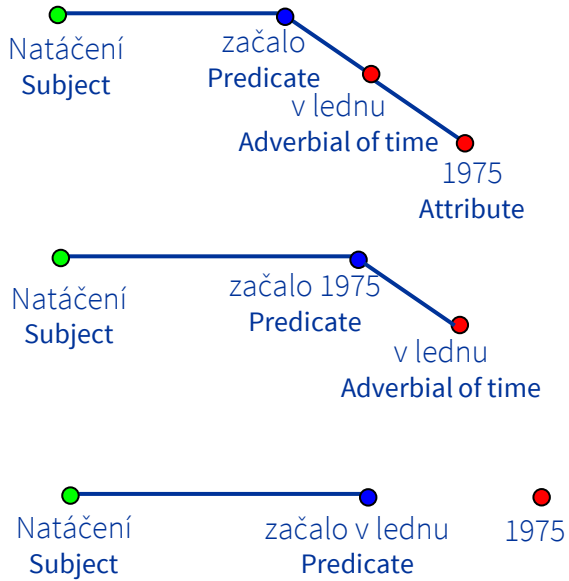
Sentence diagram over sentence $s = w_1\, w_2\, \ldots w_n$ is a directed acyclic graph $D = (Nodes, Edges)$ where $Nodes$ is a partition of s and $Edges = \{(N_1, N_2) : N_1, N_2 \in Nodes, N_1 \text{ is a child node of } N_2\}$.

Combination of diagrams

Example

Natáčení začalo v lednu 1975. Filming began in January 1975.

multiple diagrams



Data quality:

Agreement on the output

means more reliable results

→

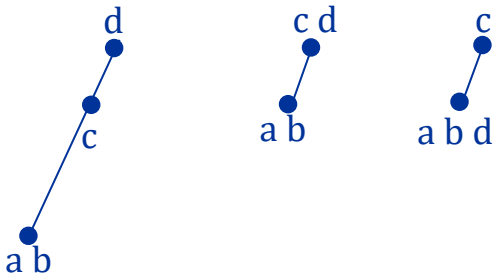
Combine diagrams by majority voting

Combination of diagrams: Nodes

Combination of $D_1, D_2, \ldots, D_m$ over sentence $s = w_1\, w_2\, \ldots w_n$

$s = a\ b\ c\ d$

D1 D2 D3



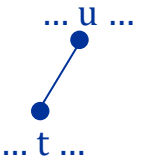
	a	b	c	d
a	x	**3**	0	1
b	x	x	0	1
c	x	x	x	1
d	x	x	x	x

$FinalNodes$

$FinalNodes = s/eq$ where $\mathrm{eq}(t,u) \leftrightarrow votes(t,u \text{ in one node}) \geq m/2$

Combination of diagrams: Edges

1. Assign weights to all word pairs (t, u) in each input diagram
2. Assign weights to all potential edges $|PotentialEdges| = \binom{|FinalNodes|}{2}$
3. Find a maximum spanning tree using (Chu and Liu, 1965), (Edmonds, 1967)

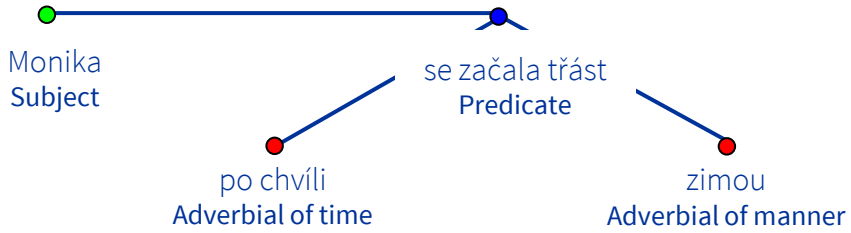



```
sort PotentialEdges                          # so that weight(E_1) ≥ weight(E_2) ≥ ⋯
FinalEdges := ∅
until PotentialEdges := ∅
    FinalEdges      := FinalEdges ∪ E_1
    PotentialEdges := PotentialEdges \ E_1
    PotentialEdges := PotentialEdges \ ¬E_1      # where ¬E_1 is reverse to E_1
    PotentialEdges := PotentialEdges \ {E: E ∪ Final Edges has a cycle}
```

Transformation of diagrams

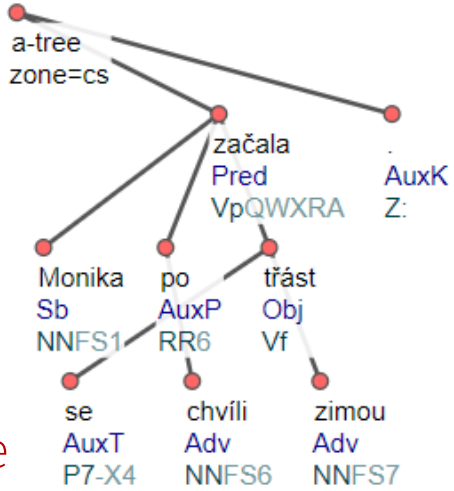
Example

Monika se po chvíli začala třást zimou. Monika started to shiver of cold in a while.



final diagram

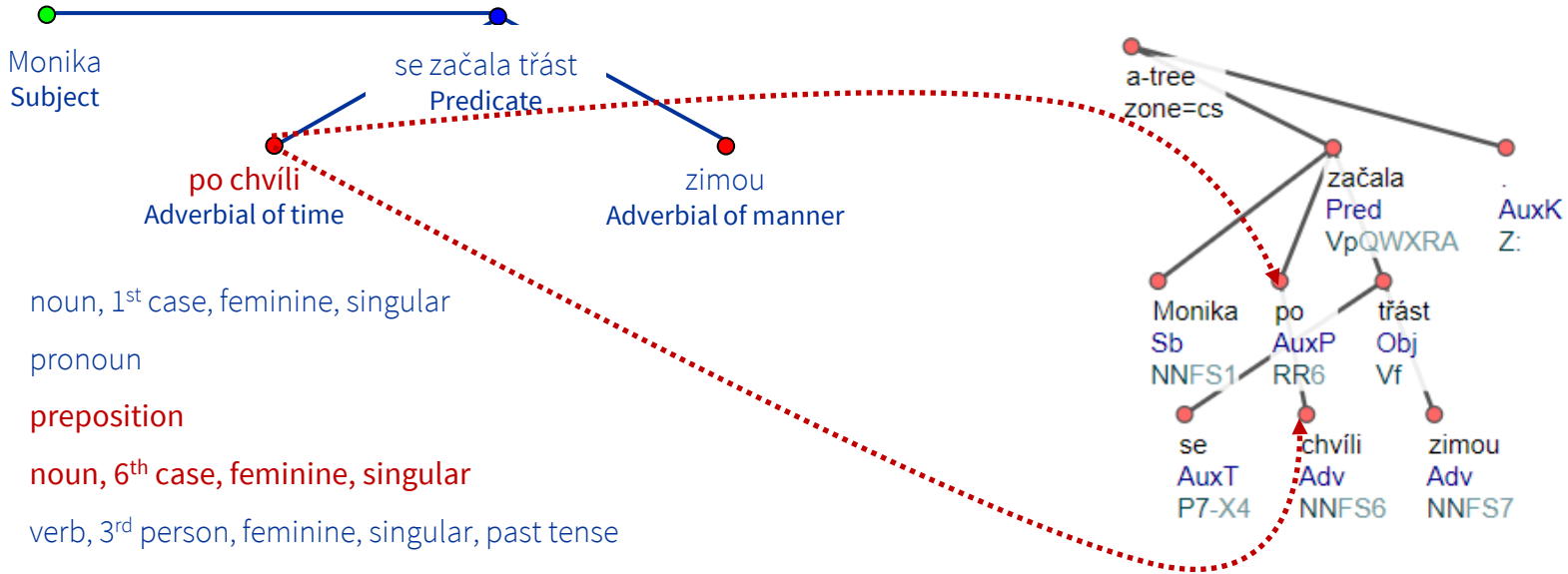
Transform it into a target scheme



Transformation of diagrams

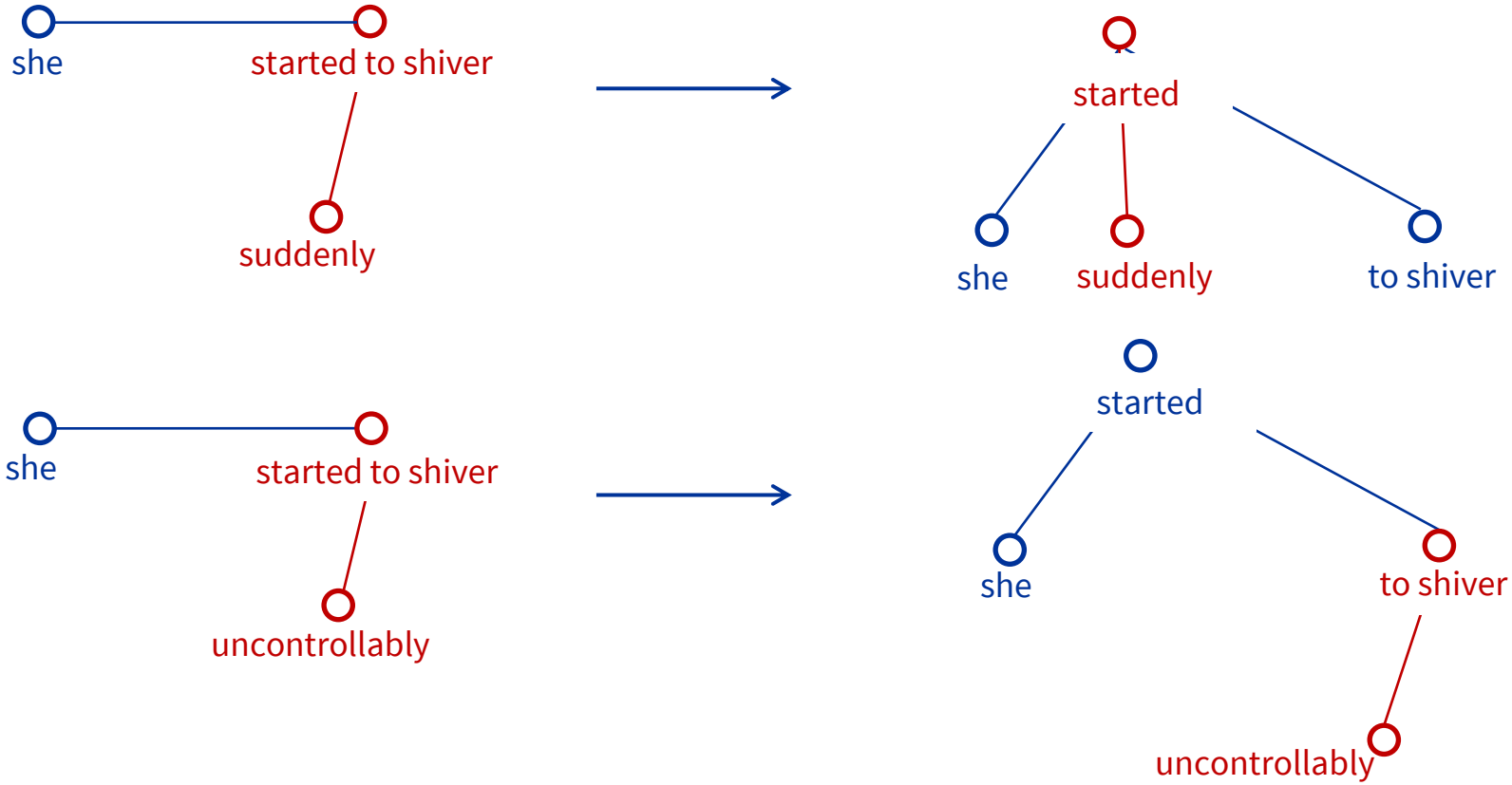
Example

Monika se po chvíli začala třást zimou. Monika started to shiver of cold in a while.



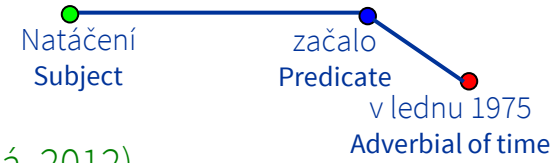
Monika	noun, 1st case, feminine, singular
se	pronoun
po	preposition
chvíli	noun, 6th case, feminine, singular
začala	verb, 3rd person, feminine, singular, past tense
třást	verb infinitive
zimou	noun, 7th case, feminine singular

Transformation of diagrams



Annotation by Crowdsourcing: Evaluation

Workbench of 100 sentences: **Sentence diagrams**



– 2 teachers and 9 students in Čapek editor (Hana, Hladká, 2012)

Similarity

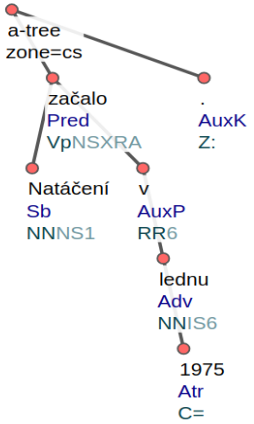
– **D**iagram **E**dit **D**istance (= the cost of the cheapest transformation between diagrams)
(Hana, Hladká, 2014)

– the smaller the distance, the more similar the diagrams
  – **DED(T_1, T_2) = 0.26**
  – DED(T_1, combination(S_3,…,S_9)) = 0.41
  – DED(S_1, S_2) = 0.69

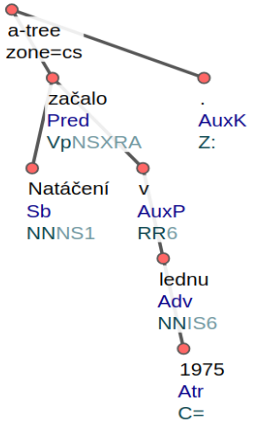
Annotation by Crowdsourcing: Evaluation

Workbench of 100 sentences: **Trees**

- correct trees by a linguist (= gold examples)
- automatically
  - statistical parser trained on Prague Dependency Treebank
  - our **combination + transformation** procedure (**ct**)



Annotation by Crowdsourcing: Evaluation



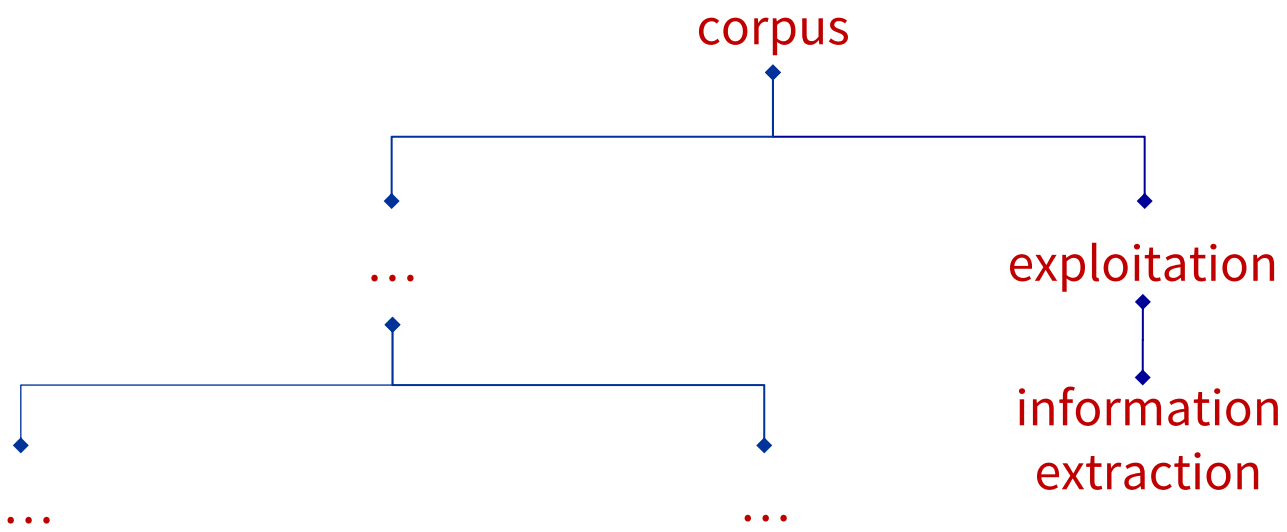
Workbench of 100 sentences: **Trees**

- correct trees by a linguist (= gold examples)
- automatically
  - statistical parser trained on Prague Dependency Treebank
  - our **combination + transformation** procedure (**ct**)

Similarity

- **U**nlabeled **A**ttachment **S**core (= the percentage of words having the correct heads)
- the higher the score, the more similar trees
- UAS(parser) = **0.9**, UAS(**ct**(T_1, T_2)) = **0.9**, UAS(**ct**(S_3,…,S_9)) = **0.77**

Topic #2



Corpus – Exploitation – Information extraction

Information extraction is a process that turns unstructured information embedded in texts into structured data.

Example

One Flew Over the Cuckoo's Nest is a 1975 American comedy-drama film directed by Miloš Forman, based on … by Ken Kesey …

Named-entity recognition, e.g. person names
Miloš Forman, Ken Kesey

Corpus – Exploitation – Information extraction

Example

One Flew Over the Cuckoo's Nest is a 1975 American comedy-drama film directed by Miloš Forman, based on … by Ken Kesey …

Relation extraction

Miloš Forman is the director of the movie One Flew Over the Cuckoo's Nest

Legal domain

- Specialized complex language
- Sentences well constructed but often long and highly structured, with embedded clauses

Example

Účetní jednotky účtují o stavu a pohybu majetku a jiných aktiv, závazků a jiných pasiv, dále o nákladech a výnosech a o výsledku hospodaření.

Accounting units shall account for the position of, and movements in, their property and other assets, commitments and other liabilities as well as costs and revenues, and their business result.

Topic #2

We are developing an extraction system over legal texts. Is the extraction over dependency trees instead of free texts beneficial to the system?

Our extraction task

Entity and relation extraction from legal texts

- Entities from the accounting domain
- Relations: right, obligation, definition
- Czech

Extraction over texts?

Example

Účetní jednotky účtují o stavu a pohybu majetku a jiných aktiv, závazků a jiných pasiv, dále o nákladech a výnosech a o výsledku hospodaření.

Accounting units shall account for the position of, and movements in, their property and other assets, commitments and other liabilities as well as costs and revenues, and their business result.

what accounting units have to do

Extraction over texts?

Example

Účetní jednotky účtují o stavu a pohybu majetku a jiných aktiv, závazků a jiných pasiv, dále o nákladech a výnosech a o výsledku hospodaření.

Účetní jednotky účtují: stav majetku$_1$, stav aktiv$_2$, stav závazků$_3$, stav pasiv$_4$, pohyb majetku$_5$, pohyb aktiv$_6$, pohyb závazků$_7$, pohyb pasiv$_8$, náklady$_9$, výnosy$_{10}$, výsledek hospodaření$_{11}$

Accounting units shall account for the position of, and movements in, their property and other assets, commitments and other liabilities as well as costs and revenues, and their business result.

Accounting units account for: position of property$_1$, position of assets$_2$, position of commitments$_3$, position of liabilities$_4$, movement in property$_5$, movement in assets$_6$, movement in commitments$_7$, movement in liabilities$_8$, costs$_9$, revenues$_{10}$, business result$_{11}$

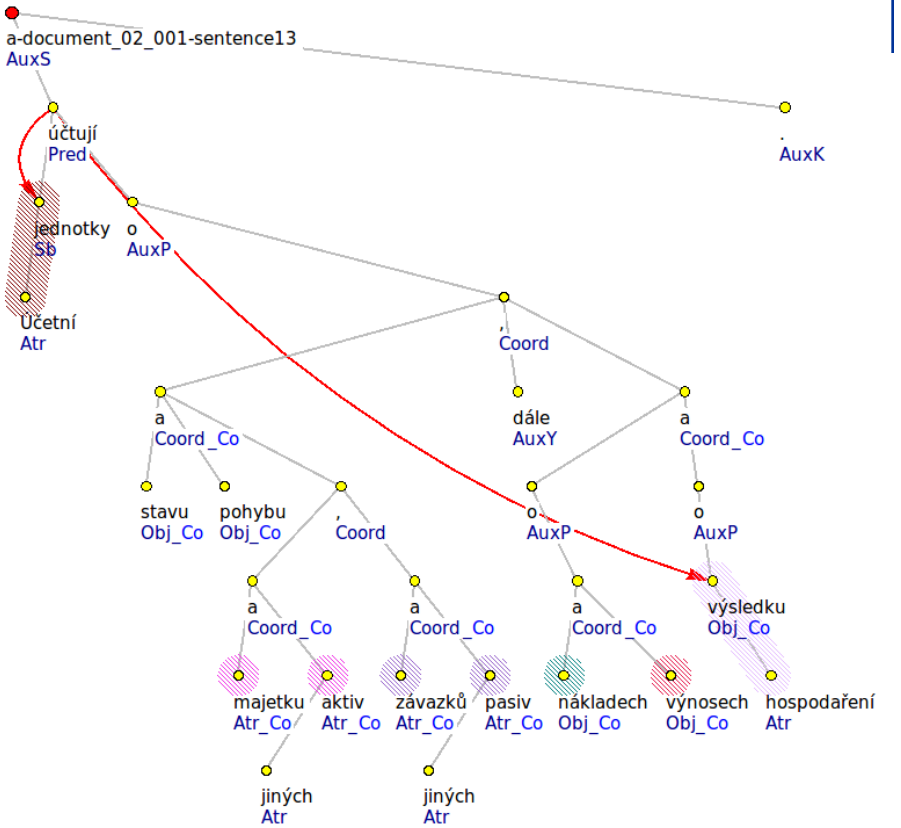
Our extraction task

~~Extraction over texts?~~

Extraction over dependency trees

- Motivated by (Mintz et al, 2009)
- Trees automatically by parser

Long-range relations between words but relatively short paths in tree



Účetní jednotky účtují o stavu a pohybu majetku a jiných aktiv, závazků a jiných pasiv, dále o nákladech a výnosech a o výsledku hospodaření.
Accounting units shall account for the position of, and movements in, their property and other assets, commitments and other liabilities as well as costs and revenues, and their business result.

Corpus of legal texts

Academic annotation to evaluate accuracy of parser and extraction procedure

- Czech Legal Text Treebank 1.0, 35K words, morph+synt (Kríž, Hladká, Urešová, 2016)
- Czech Legal Text Treebank 2.0, morph+synt+entities+relations (Kríž, Hladká, 2018)

RExtractor extraction procedure (Kríž, Hladká, 2015)

Input	text
Output	triple (subject-entity, predicate, object)

Example

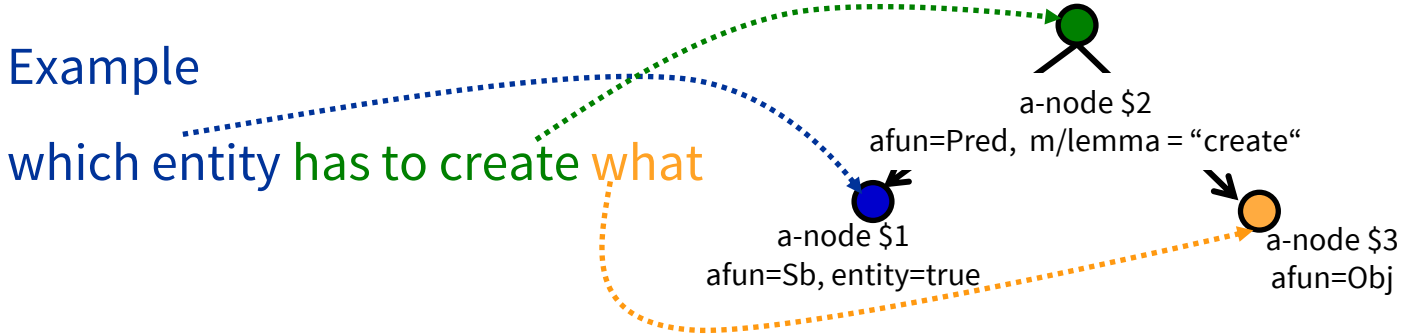
Input	**(3) Účetní jednotky jsou povinny inventarizovat majetek a závazky podle §29 a 30.**
Output	(účetní jednotky, inventarizovat, majetek), (účetní jednotky, inventarizovat, závazky)

Input	**(3) Accounting units shall take inventory of their assets and liabilities pursuant …**
Output	(accounting units, take inventory, assets), (accounting units, take inventory, liabilities)

RExtractor

- Gazetteer of accounting entities
- Gazetteer of lexical expressions for obligation, right, and definition
- Triples by querying dependency trees
  - Entities and relations as tree queries (patterns)
  - PML Tree Query language (Pajas, Štěpánek, 2009)

Example

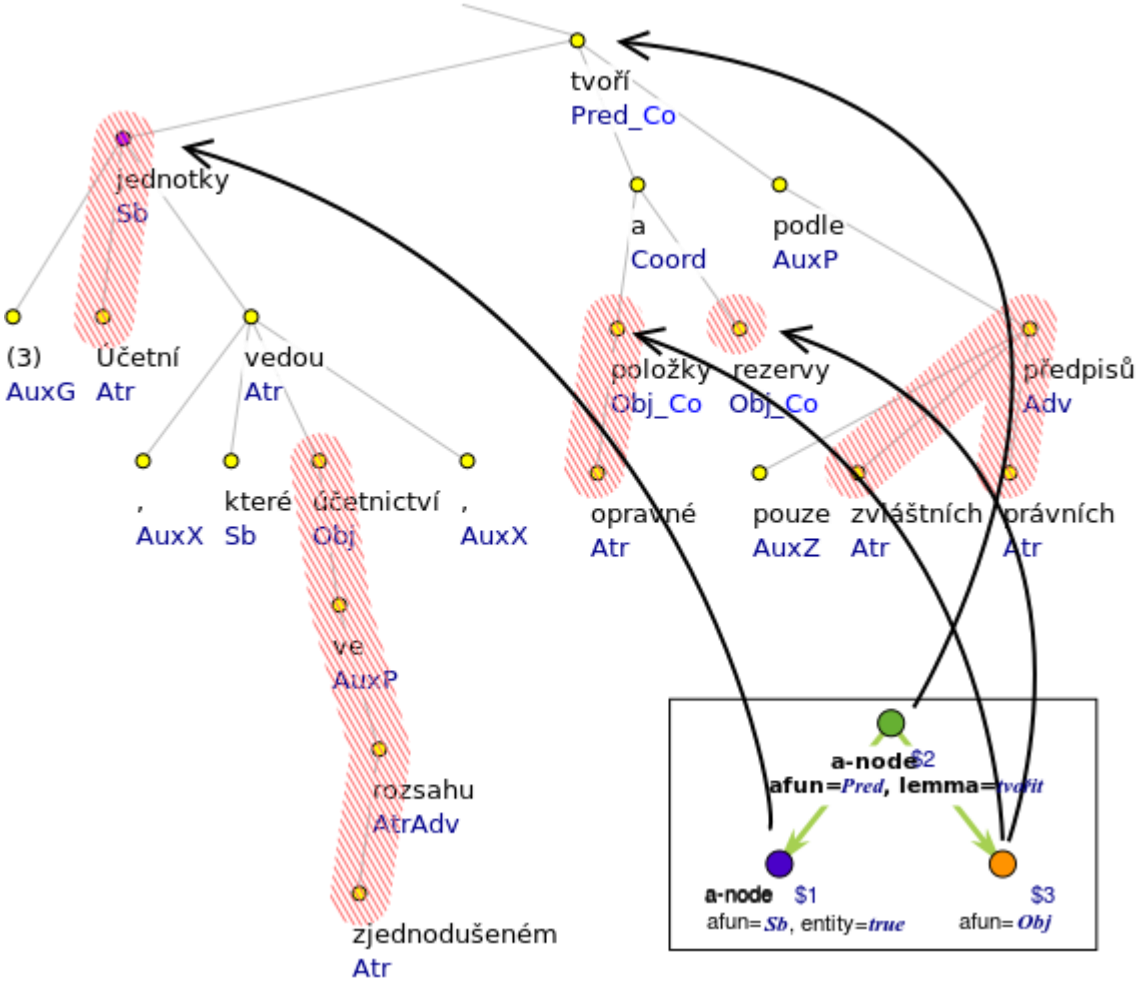


RExtractor

Example

(3) Účetní jednotky, …, tvoří opravné položky a rezervy …

(3) Accounting units, …, create fixed items and reserves …

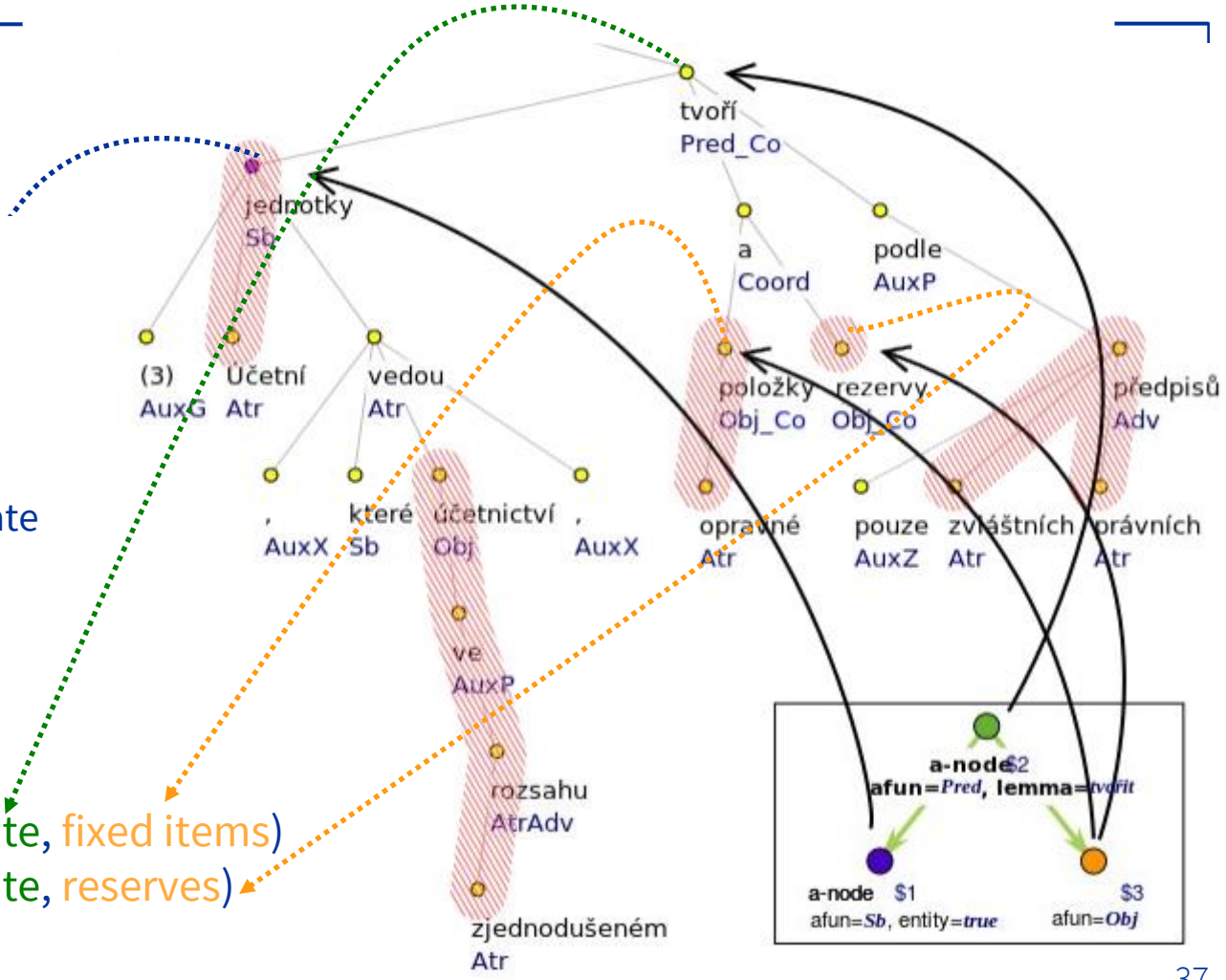


RExtractor

Example

(3) Účetní jednotky, …, tvoří opravné položky a rezervy …

(3) Accounting units, …, create fixed items and reserves …



(accounting units, create, fixed items)
(accounting units, create, reserves)

Improving parsing

- Parsers for Czech trained on newspaper texts
- Cross-domain parsing adaptation

  newspaper texts > legal texts

Linguistic analysis of legal and newspaper texts

(Kríž, Hladká, Urešová, 2016)

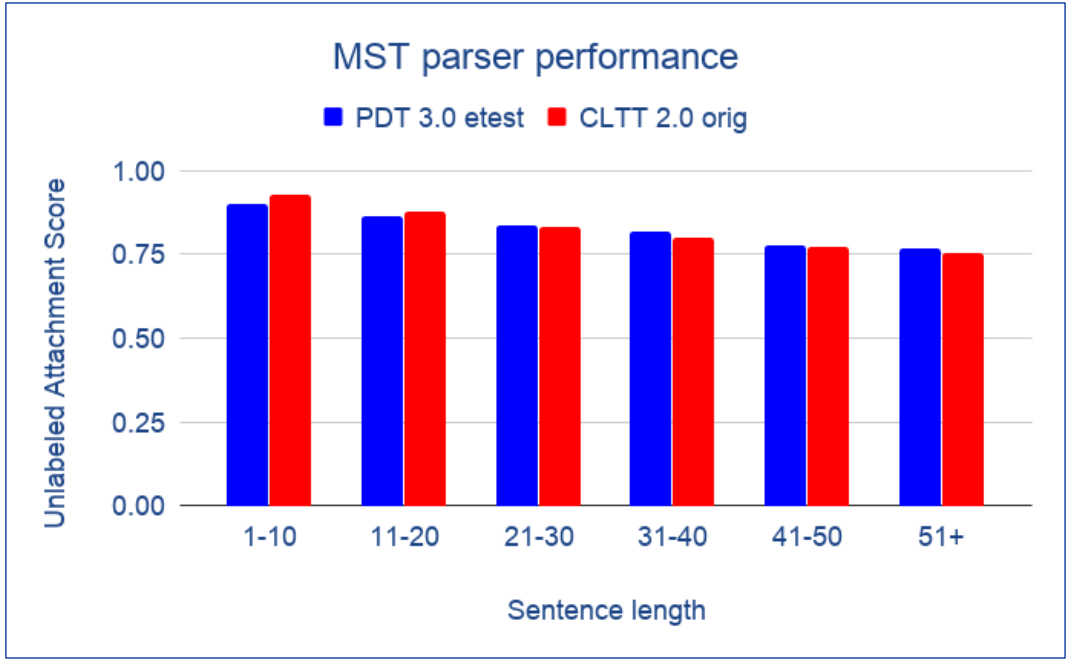
		linguistic phenomena (%)			
corpus	ASL	1.	2.	3.	…
Czech Legal Text Treebank 1.0	31.0	41.67	44.68	20.92	…
Prague Dependency Treebank 3.0	16.9	9.56	25.61	8.09	…

Average Sentence Length (ASL), **1.** chains of two genitive expressions, **2.** ellipsis, **3.** parenthesis,

Example

In the legal texts, 41.67% of the nodes are the heads of genitive phrases, like **shromažďování záznamů** (**gathering records**)

Performance of MST parser

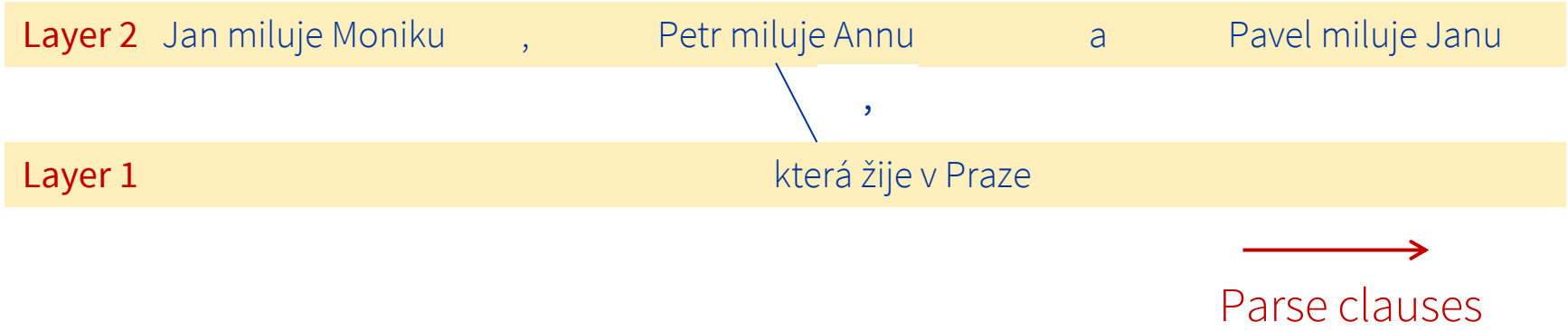


Segment sentences into clauses

Split-Parse-Join method (Kríž, Hladká, 2016)

Example

Jan miluje Moniku, Petr miluje Annu, která žije v Praze, a Pavel miluje Janu.
John loves Monica, Peter loves Ann, who lives in Prague, and Paul loves Jane.

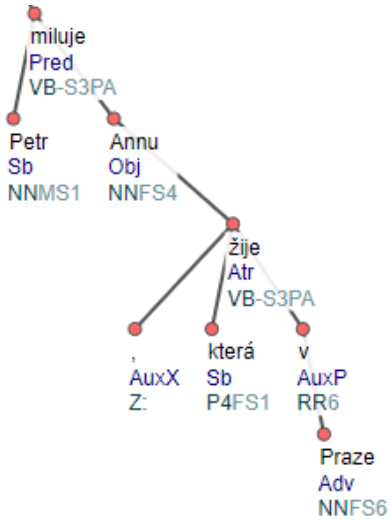


Split-Parse-Join method

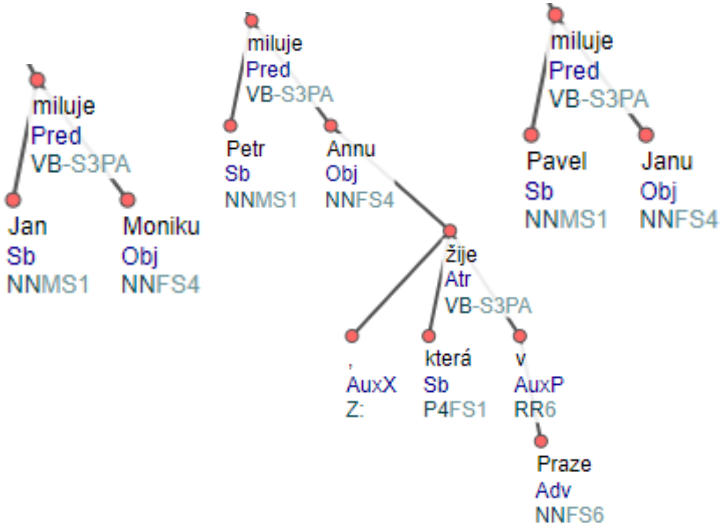
1. Parse neighboring coordinated clauses at the lowest layer separately
2. Parse the longest sequence of subordinated and coordinated clauses at the neighboring layers
3. Eliminate at least one layer

Split-Parse-Join method

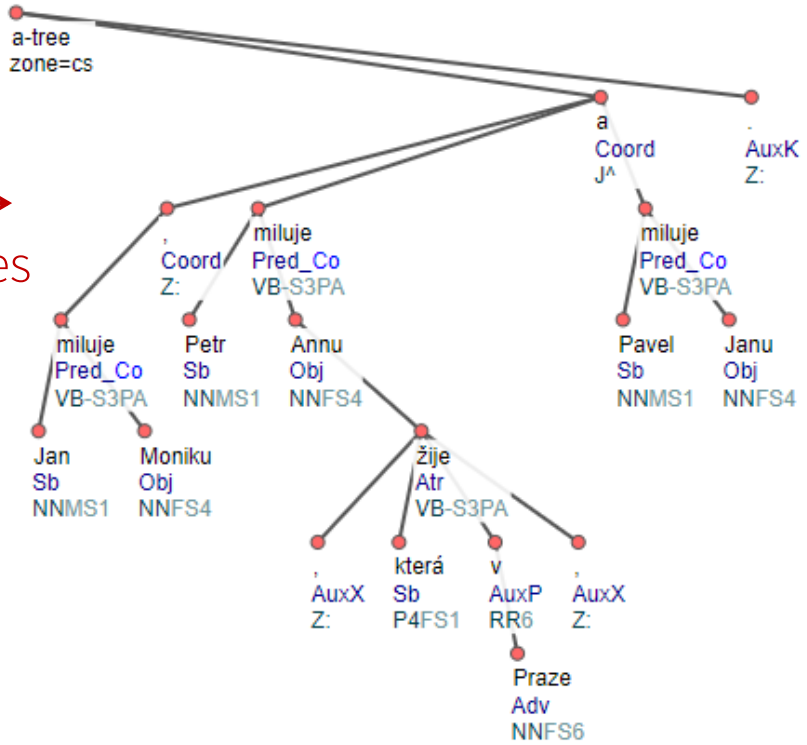
Layer 1 Jan miluje Moniku , a Pavel miluje Janu



Split-Parse-Join method

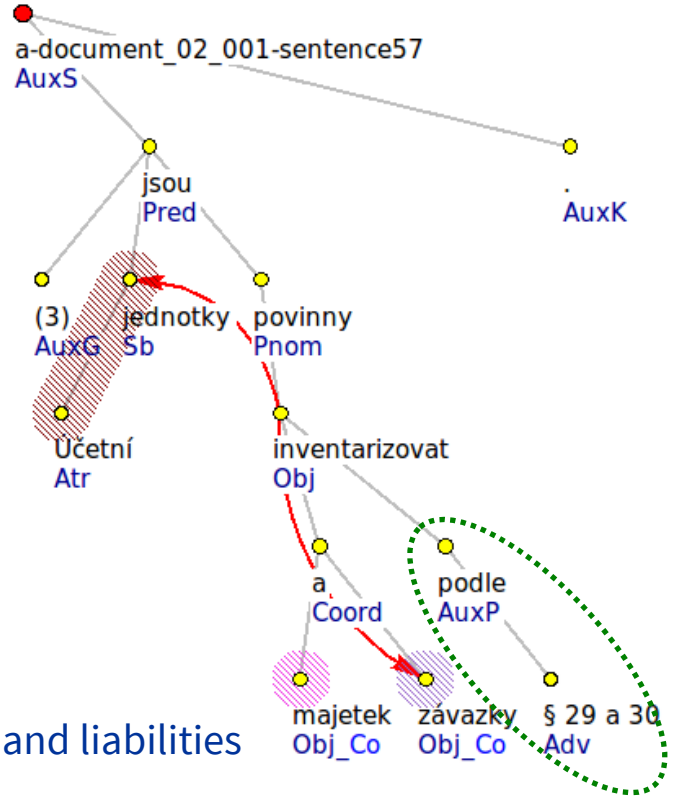


Join trees



Information extraction over trees: Evaluation

- Gold examples CLTT2.0 test
- Precision(full text search) = 0.54
- Recall(full text search) = 0.47
- **Precision(RExtractor) = 0.77**
- **Recall(RExtractor) = 0.68**
- More information
  in the subtrees of the triple's members



Example

(3) Accounting units shall take inventory of their assets and liabilities pursuant to section 29 and 30.

My contribution

- to a substantial increase of the annotated data volume
- to a formulation of alternative annotation strategy to decrease annotation costs
- to an exploration of annotated data to use them as a grammar workbook
- to an exploration of understudied domains
- to a study of parsing procedure and its potential mainly for complex sentence parsing
- to a design of extraction system over dependency trees

Academic annotation

Prague Dependency Treebank

- Böhmová A., Hajič J., Hajičová E., **Hladká B**. The Prague Dependency Treebank: A Three-Level Annotation Scenario. In: Abeillé A. (eds) I, vol 20. Springer, Dordrecht, 2003. 465 citations in Google schoolar as of May 30, 2020

Czech Academic Corpus [PI: GA AV ČR 1ET101120413, 2004-2008]

- **Hladká B.,** Hajič J., Hana J., Hlaváčová J., Mírovský J., Votrubec J.: *Czech Academic Corpus 1.0 Guide.* Karolinum, Charles University Press, Prague, Czech Republic, ISBN 978-80-246-1315-4, 100 pp., 2007. scientific monograph
- **Hladká B.**, Bémová A., Urešová Z.: Syntaktická proměna Českého akademického korpusu. *Slovo a slovesnost*, Vol. 72, No. 4, Ústav pro jazyk český AV ČR, ISSN 0037-7031, pp. 268-287, 2011. IF in 2018: 0.412
- **Hladká B.**, Králík J.: Proměny Českého akademického korpusu. *Slovo a slovesnost*, Vol. 67, No. 4, Ústav pro jazyk český AV ČR, ISSN 0037-7031, pp. 179-194, 2006. IF in 2018: 0.412

Data collections

1. Hajič J., Hajičová E., Pajas P., Panevová J., Sgall P., **Vidová Hladká B**. *Prague Dependency Treebank 1.0,* LDC - Linguistic Data Consortium, 2001. (url)
2. Hana J., **Hladká B.**: *CzeSL - Universal Dependencies Release 0.5*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic, 2019. (url)
3. **Hladká B.**, Hajič J., Hana J., Hlaváčová J, Mírovský J., Raab J.: *Czech Academic Corpus 2.0*, LDC - Linguistic Data Consortium, 2008. (url)
4. **Hladká B.**, Kučera O., Kuchyňová K.: *STYX 1.0*. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic, 2017. (url)
5. Kríž V., **Hladká B.** *Czech Legal Text Treebank 2.0*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic, 2017. (url)
6. Kríž V., **Hladká, B.,** Urešová, Z.: *Czech Legal Text Treebank 1.0*, LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic, 2015. (url)
7. Kríž V., **Hladká B.**: *Czech Court Decisions Dataset,* LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic, 2014. (url)
8. Nivre J., Abrams M., Agić Ž. et al: *Universal Dependencies 2.3*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, 2018. (url)

Alternative annotation by Crowdsourcing

- **Hladká B.**, Mírovský J., Schlesinger P.: Play the Language: Play Coreference. In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, Association for Computational Linguistics, Suntec, Singapore, ISBN 978-1-932432-61-9, pp. 209-212, 2009. CORE ACL – A*, CORE IJCNLP – B
- **Hladká B.**, Kučera O.: Prague Dependency Treebank as an Exercise Book of Czech. In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Vancouver, BC, Canada, ISBN 1-932432-55-8, pp. 14-15, 2005. CORE-A

Information extraction [TAČR TA02010182 INTLIB, 2012-2015]

- Kríž V., **Hladká B.**: Improving Dependency Parsing Using Sentence Clause Charts. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* – Student Research Workshop, Association for Computational Linguistics, Stroudsburg, PA, USA, ISBN 978-1-945626-02-9, pp. 86-92, 2016. CORE-A
- Kríž V., **Hladká B.**: RExtractor: a Robust Information Extractor. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, Association for Computational Linguistics, Denver, CO, USA, ISBN 978-1-941643-49-5, pp. 21-25, 2015. CORE-A
- Kríž V., **Hladká B.**, Nečaský M., Knap T.: Data Extraction Using NLP Techniques and Its Transformation to Linked Data. In: 13th Mexican International Conference on Artificial Intelligence, MICAI 2014, Tuxtla Gutiérrez, Mexico, November 16-22, 2014. Proceedings, Part I, Copyright © Springer International Publishing, Switzerland, ISBN 978-3-319-13646-2, pp. 113-124, 2014.

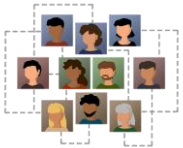
Future plans

- **Education**
  - Čapek development towards learning application (data-driven)
  - More annotation schemes, non-native speakers (Hana, Hladká, 2018), (Hladká, Holub, Kríž, 2013), (Ircing, Švec, Zajíc, Hladká, Holub, 2017)

- **Legal domain processing**
  - Apply RExtractor to other legal subdomains and document types
  - Get real-user evaluation of Rextractor (Plaňavová Latanowicz, 2019)
  - Study clarity of legal documents over dependency trees

Acknowledgement

Alla Bémová, Alena Chrastová, Jan Hajič, Eva Hajičová, Jirka Hana, Martin Holub, Fred Jelinek, Vladislav Kuboň, Markéta Lopatková, Jiří Mírovský, Jarmila Panevová, Nino Peterek, Kiril Ribarov, Petr Sgall, Zdeňka Urešová, Dan Zeman



Pavel Ircing, Martin Nečaský, Jana Plaňavová-Latanowicz, František Plášil

Marie Konárová, Vincent Kríž, Ondřej Kučera, Karolína Kuchyňová, Bohdan Maslowski, Ivana Sixtová

Nicoleta Calzolari, Annete Frank, Pavel Smrž

Jaroslav Pokorný, Tomaž Erjavec, Luděk Műller, Joakim Nivre, Pavel Pecina

my family