Detection of attribution in Czech news

Barbora Hladká hladka@ufal.mff.cuni.cz
Radek Mařík marikr@fel.cvut.cz

in cooperation with Matyáš Kopp, Jiří Liška, Jiří Mírovský

Automation of sources in journalistic discourse - examples of good practice, October 27, 2022

Task

Automatically detect sources
that journalists credit in newspaper stories.

attribution = source + information + signal

According to NetMonitor, there are 7.77 million Internet users over the age of ten in the Czech Republic.

Data

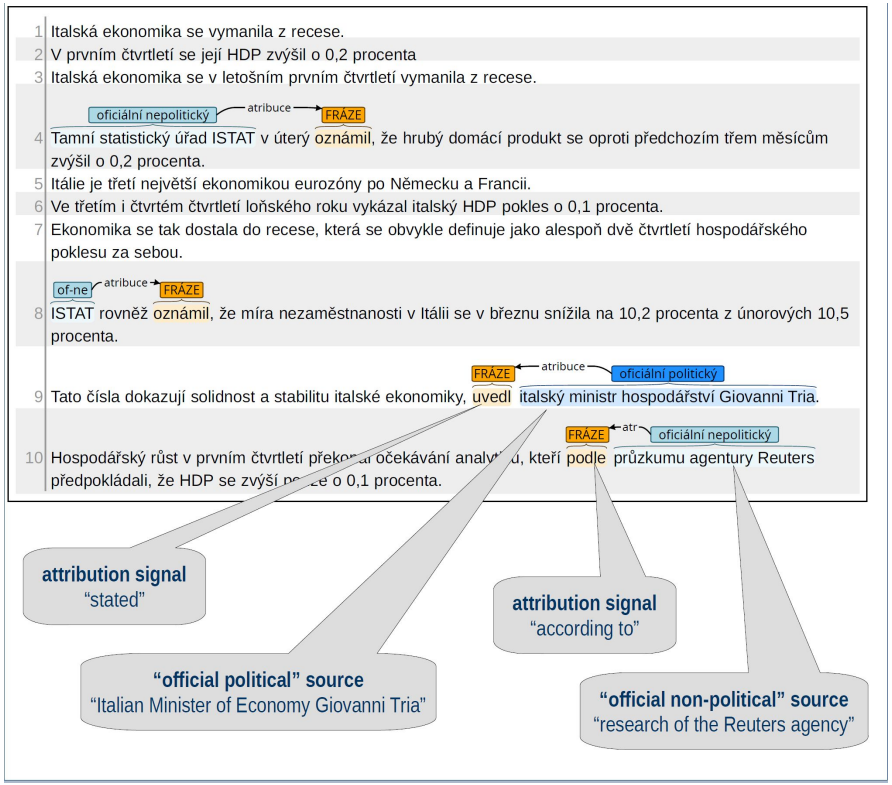
iRozhlas collection

- iRozhlas news server of the Czech public radio
  https://www.irozhlas.cz/
- 104,678 Czech articles
- period: 2017-04-01 – 2022-05-31

SiR 1.0

- manual detection of sources + signals
  in the subset of iRozhlas collection
- crowdsourcing task (https://ufal.mff.cuni.cz/anotace-citacnich-frazi-v-datech-irozhlas)
  - output: SiR 1.0 (**S**ource in **iR**ozhlas)
    = 1,718 articles (42,890 sentences, 614,995 words)

SiR 1.0 :: Crowdsourcing task



- a group of 290 students
- Brat editor
- triple annotation
  - 46 articles
  - three annotators + arbiter
- double annotation
  - 543 articles
  - two annotators + automatic unification
- single annotation
  - 1,129 articles
  - single annotator

SiR 1.0 :: Crowdsourcing task

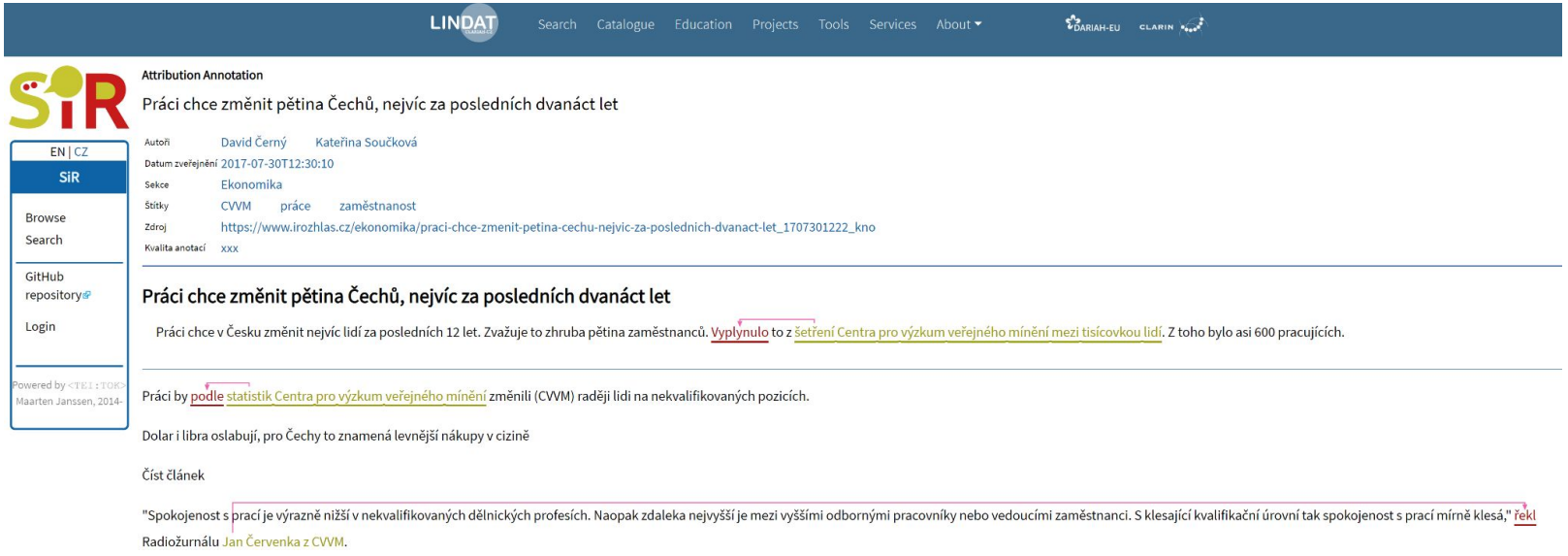
Inter-Annotator Agreement

Inter-annotator agreement in recognition of signals, sources and source classes by two annotators; measured on 170 documents.

annotation	measure	agreement
signals	F1	0.67
sources	F1	0.60
source classes	%	74
source classes	κ	0,58

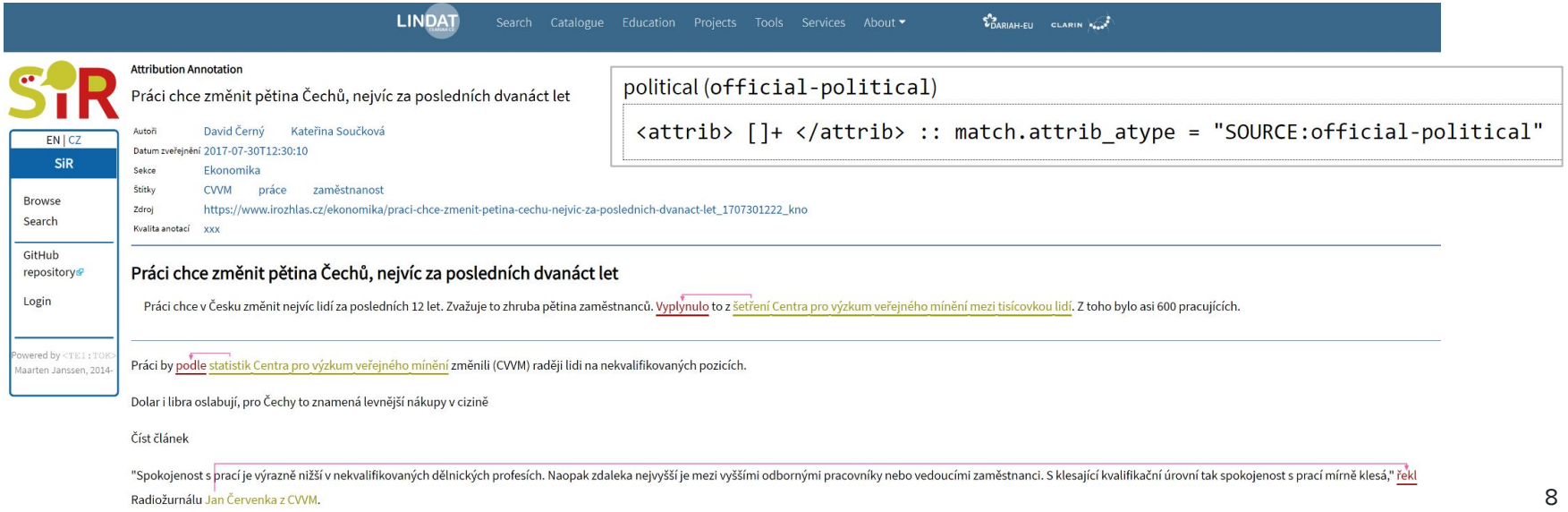
SiR 1.0

- Download from LINDAT repository http://hdl.handle.net/11234/1-4840
- View & Search in TEITOK
  https://quest.ms.mff.cuni.cz/parczech/teitok/sir/en/index.php?action=home

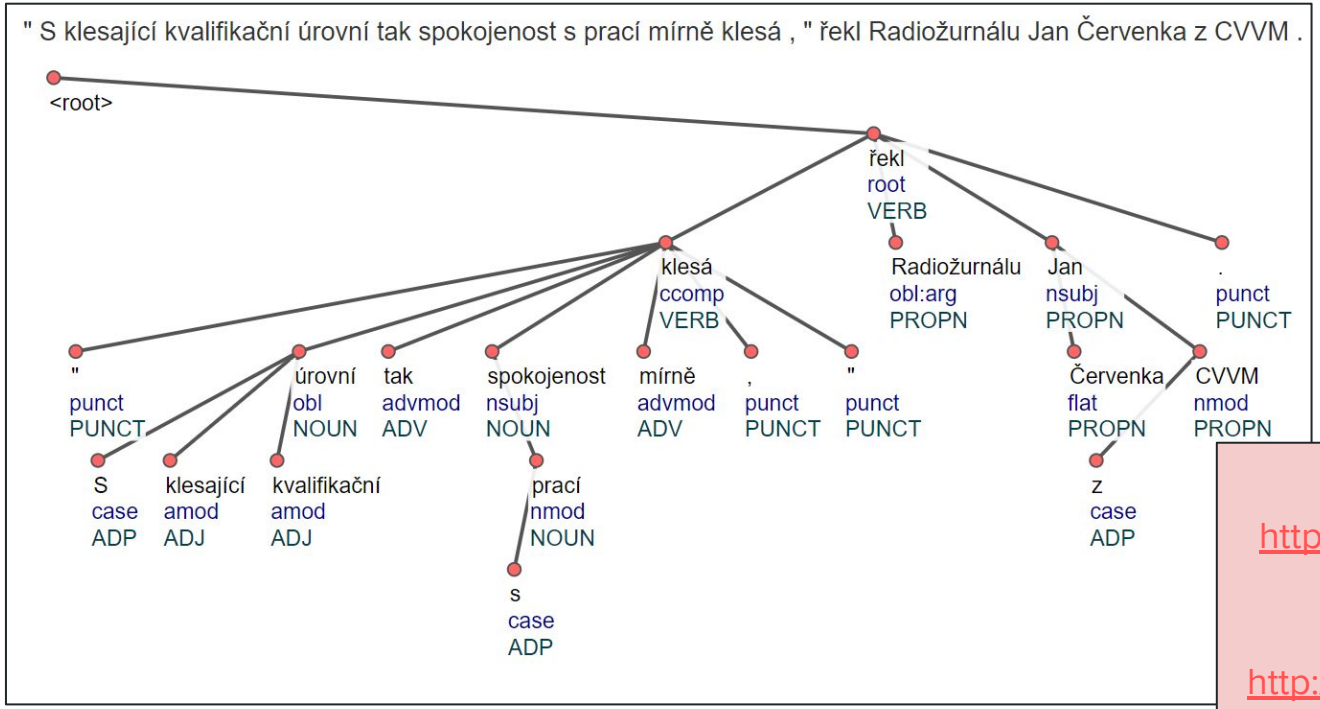


SiR 1.0

- Download from LINDAT repository http://hdl.handle.net/11234/1-4840
- View & Search in TEITOK
  https://quest.ms.mff.cuni.cz/parczech/teitok/sir/en/index.php?action=home



Automatic detection :: linguistic preprocessing



UDPipe
https://lindat.cz/services/udpipe/

NameTag
http://lindat.cz/services/nametag/

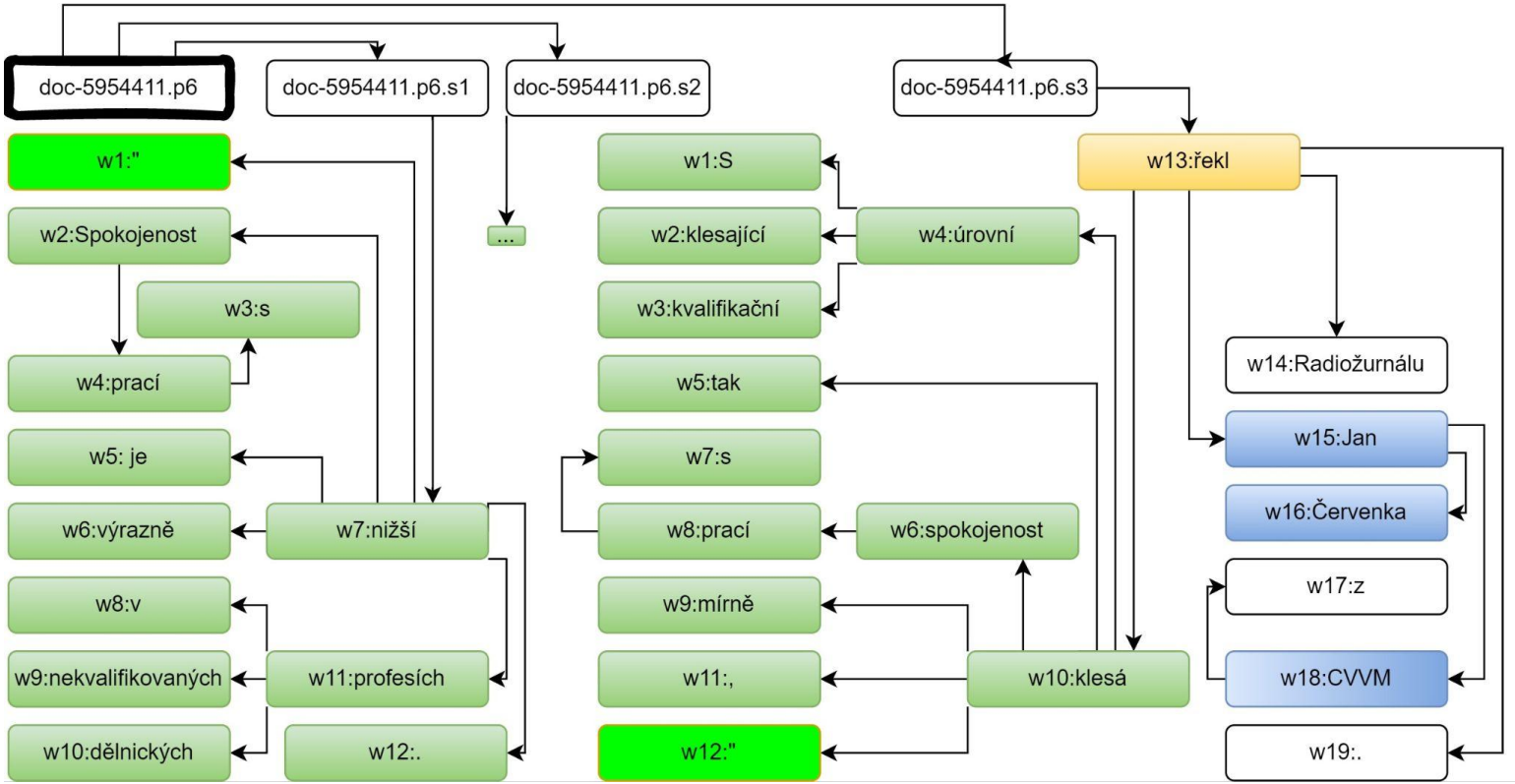
"S klesající klasifikační úrovní tak spokojenost s prací mírně klesá," řekl **Radiožurnálu** **Jan Červenka** z **CVVM**.

Automatic detection sources + signals + information

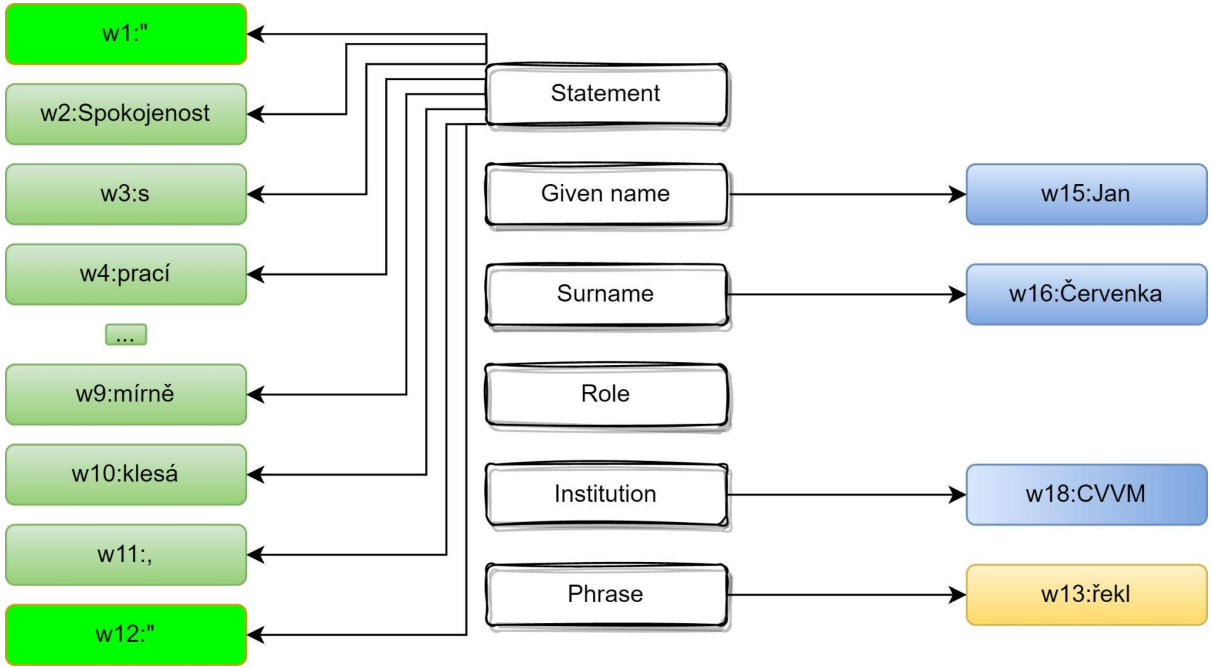
- 104,677 articles, 208,290 detected quotes
- Steps (a graph transformation):
  a. Syntactic trees using UDPipe
  b. Named-Entities using NameTag
  c. Relation filtration (ud-syn:)
     nmod, amod, flat, appos, dep
  d. Entity type filtration (ne:)
     - name - pf, ps, pm, P
     - institution - I, if, io, ic, ia, gl
     - media - ms, mn
     - geography - gu, gc, gr, G, gq, gt

- Result
  = 5-tuple + attributes (gender, etc.)
  a. Statement
     (in quotes, indirect for selected verbs)
  b. Name = given names + surname
  c. Role
  d. Institution
  e. Phrase (verb)
- Represented by lists of detected words

Automatic detection sources + signals + information

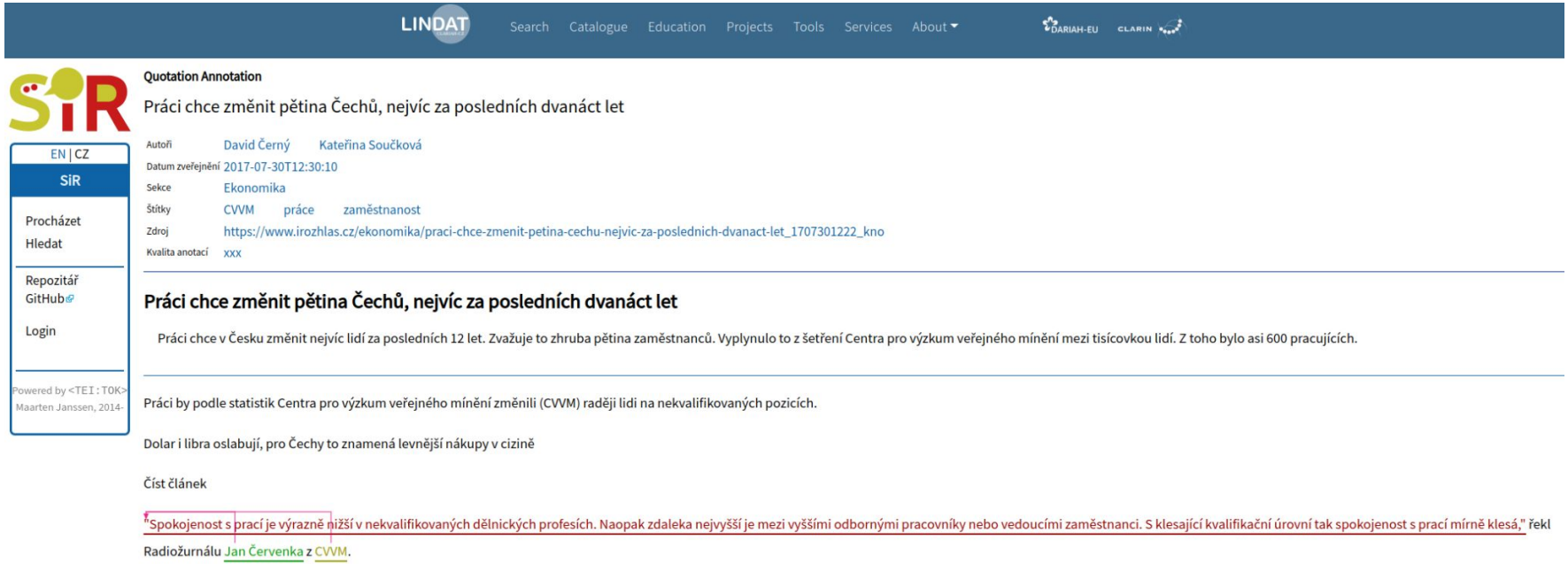


Automatic detection sources + signals + information



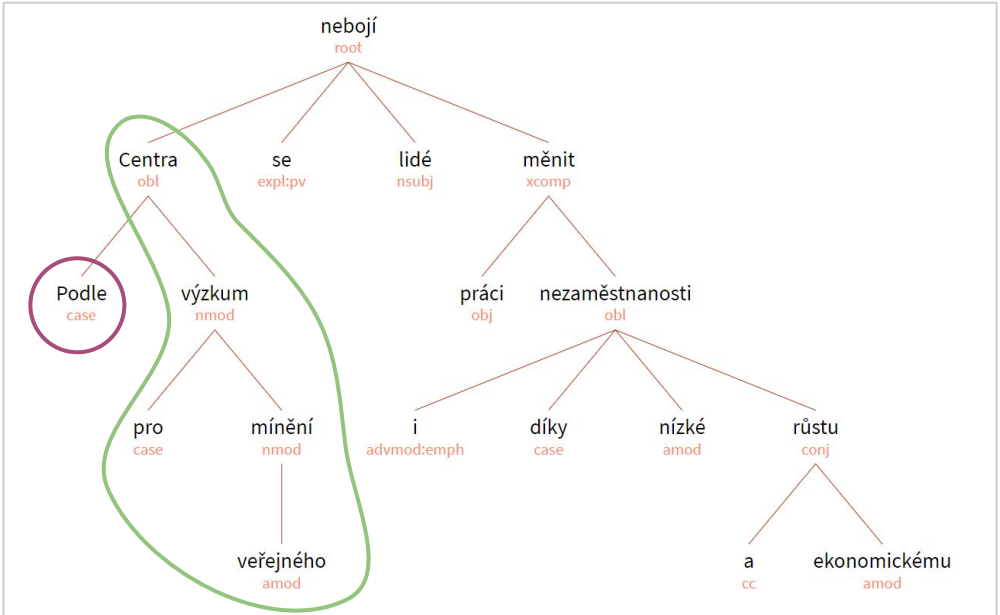
Automatic detection sources + signals + information

In TEITOK



Future plans

data mining + querying syntactic trees + SiR 1.0 list of signals



Podle Centra pro výzkum veřejného mínění se lidé nebojí měnit práci i díky nízké nezaměstnanosti a ekonomickému růstu.

References

- Hladká Barbora, Jiří Mírovský, Matyáš Kopp, Václav Moravec. Annotating Attribution in Czech News Server Articles. In: *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 1817–1823, Marseille, France 20-25 June 2022. [pdf]

Acknowledgement

- This work was supported by the Technological Agency of the Czech Republic (grant number TL05000057). This work has been using language resources and tools developed and/or stored and/or distributed by the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2018101).