Vincent Kríž, Barbora Hladká

RExtractor

Entity Relation Extraction from Unstructured Texts

Intelligent library (INTLIB, TA02010182)

Seminar of formal linguistics, 2014-05-12

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University in Prague
Czech Republic

{kriz,hladka}@ufal.mff.cuni.cz
http://ufal.mff.cuni.cz/intlib

Motivation

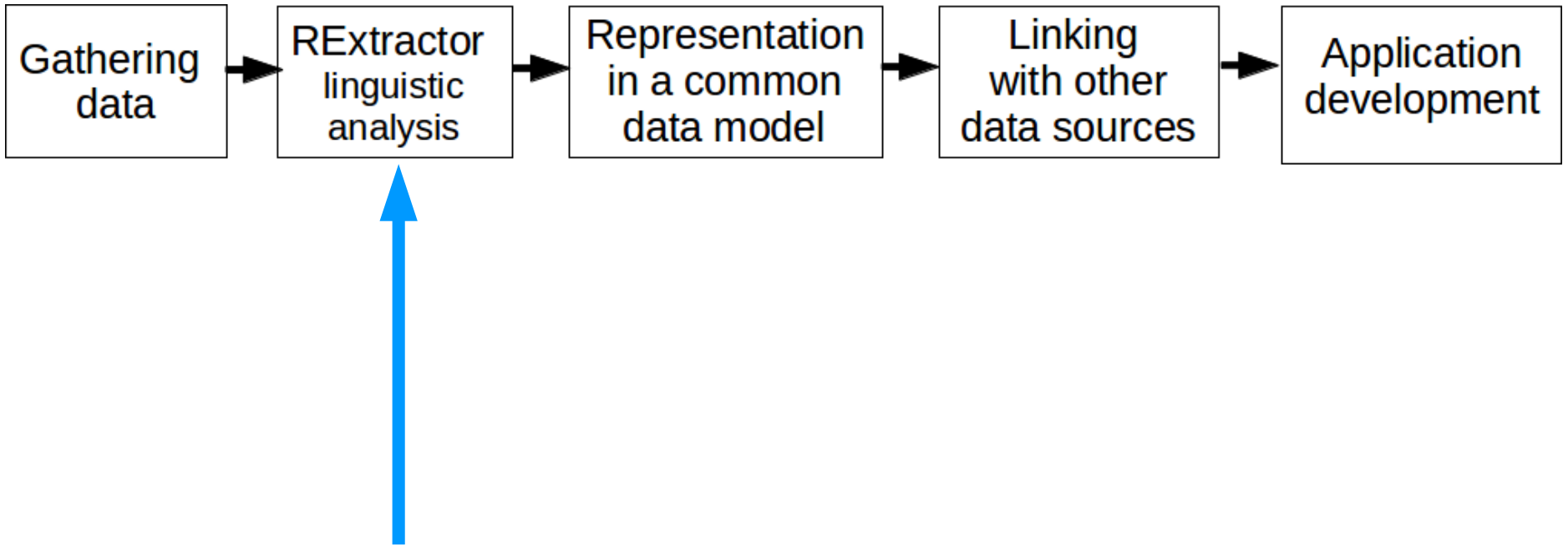
Typical search approaches

- full-text search
- metadata search

Our approach

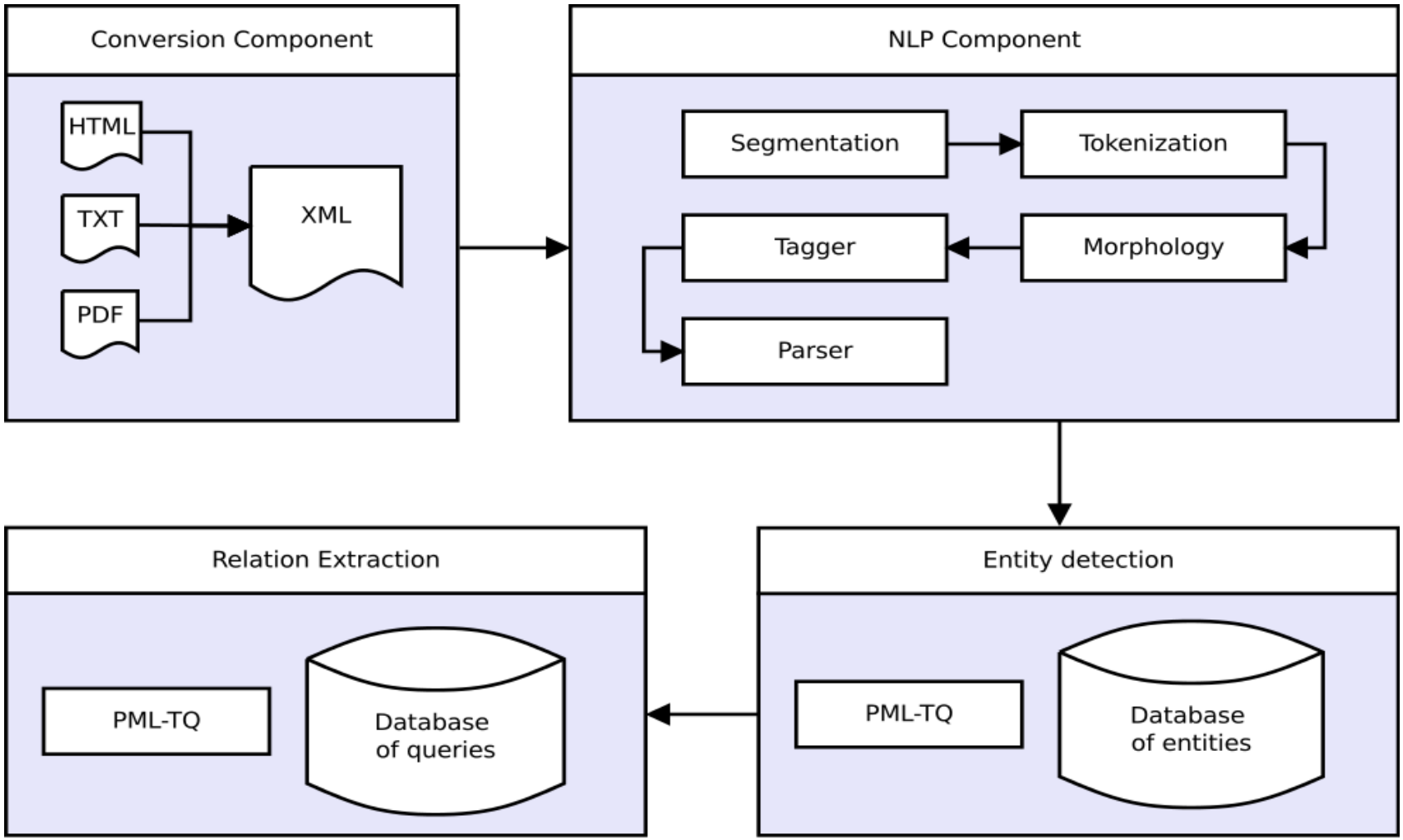
- building a knowledge base
- semantic representation of documents
- entities and their relations
- represented in the Resource Description Framework (RDF)

Data processing workflow



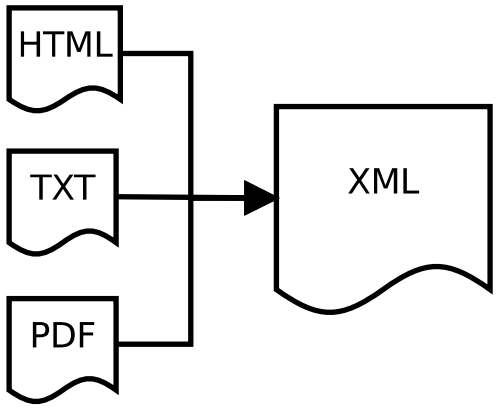
RExtractor Architecture

- Domain independent



Conversion Component

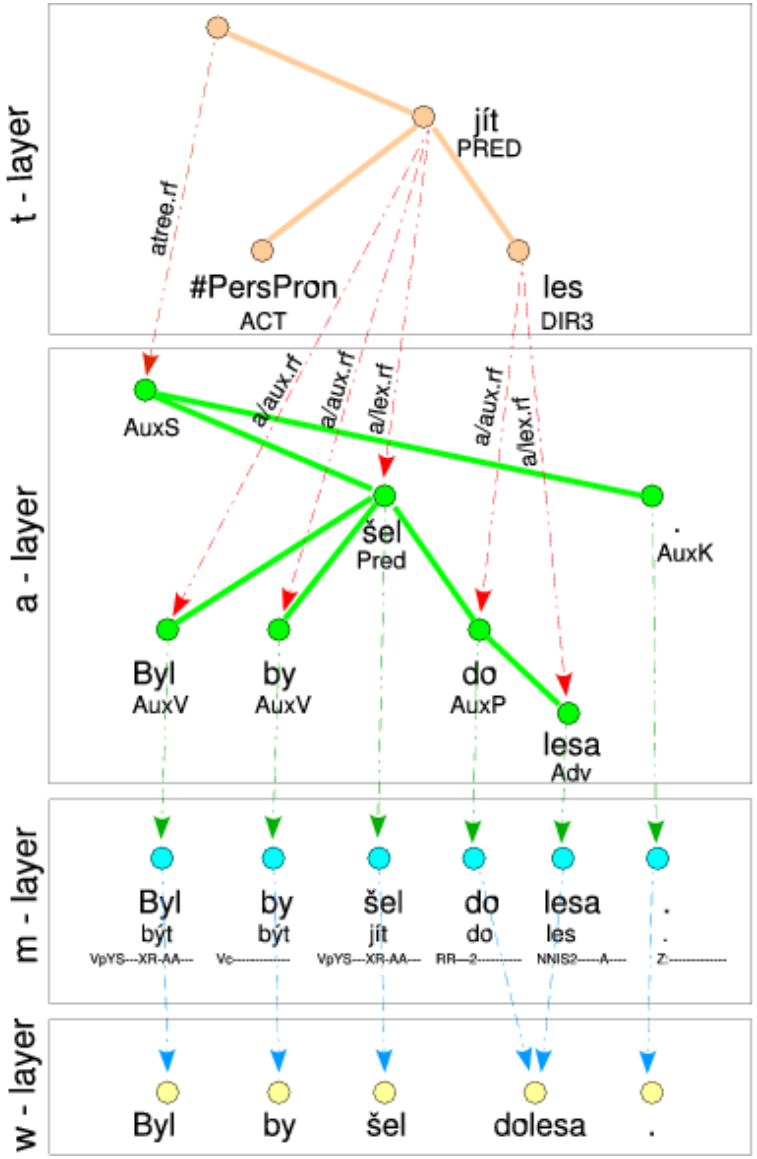
- converts various input formats into a unified representation (XML)



NLP Component

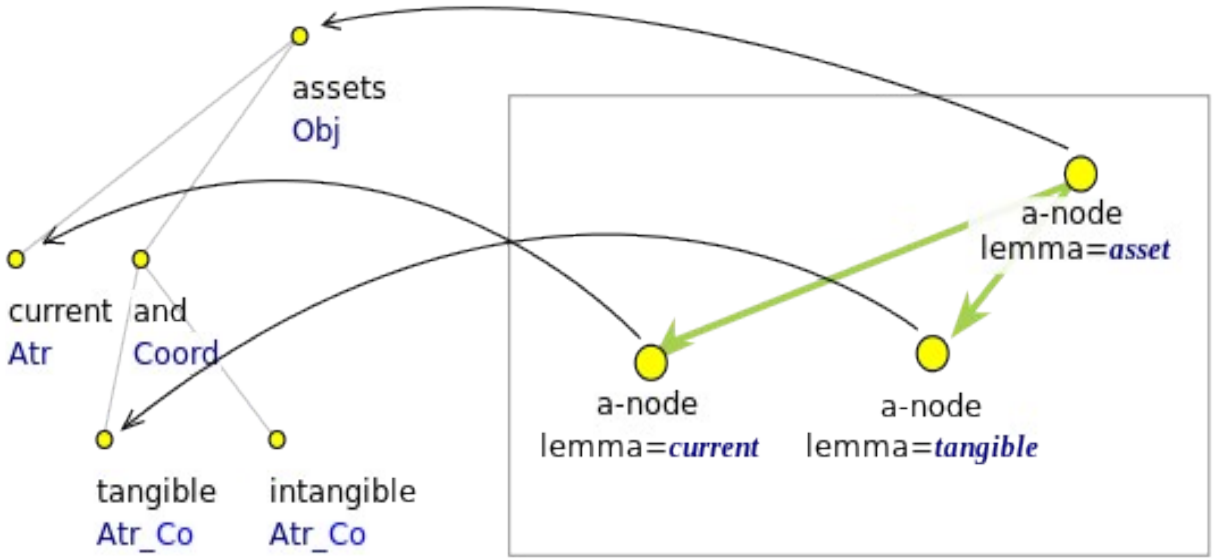
- Prague Dependency Treebank framework
- Tools
  - segmentation & tokenization
  - lemmatization & morphology
  - syntactic parsing
  - deep syntactic parsing
  - Treex

- http://ufal.mff.cuni.cz/pdt3.0
- http://ufal.mff.cuni.cz/treex



Entity Detection Component

- Database of Entities
  - entities specified by domain experts
- PML-TQ



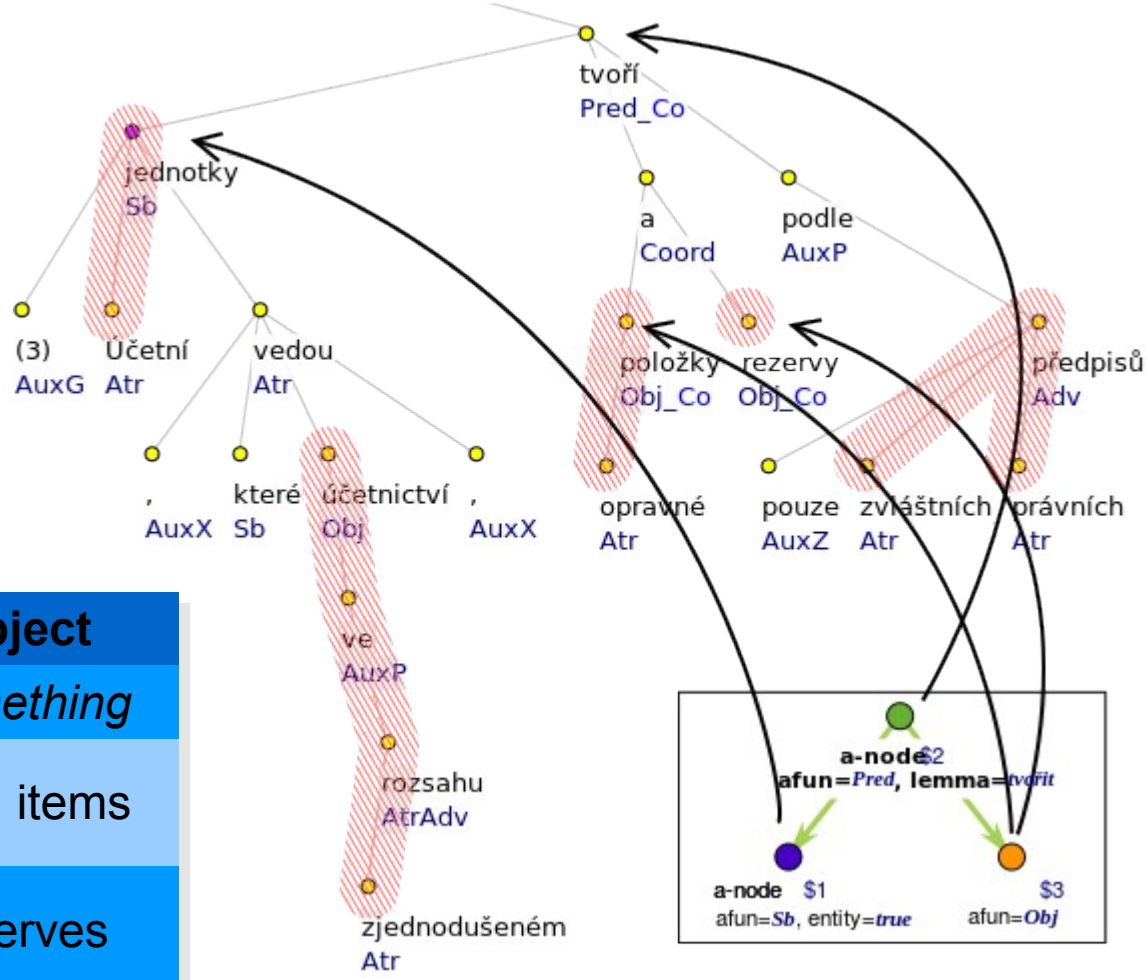
- http://ufal.mff.cuni.cz/tools/pml-tq

Relation Extraction Component

- Database of Queries
  - queries formulated by domain experts
  - their formulation in the form of PML-TQ queries on dependency trees
- RDF ready output:

Subject	Predicate	Object
Entity	*hasToCreate*	*Something*
Accounting units	create	fixed items
Accounting units	create	reserves

- Example of user query: accounting units' obligations



Case study on legislative domain

Case study on legislative domain

Legal texts

- specialized texts operating in legal settings
- they should transmit legal norms to their recipients
- they need to be clear, explicit and precise

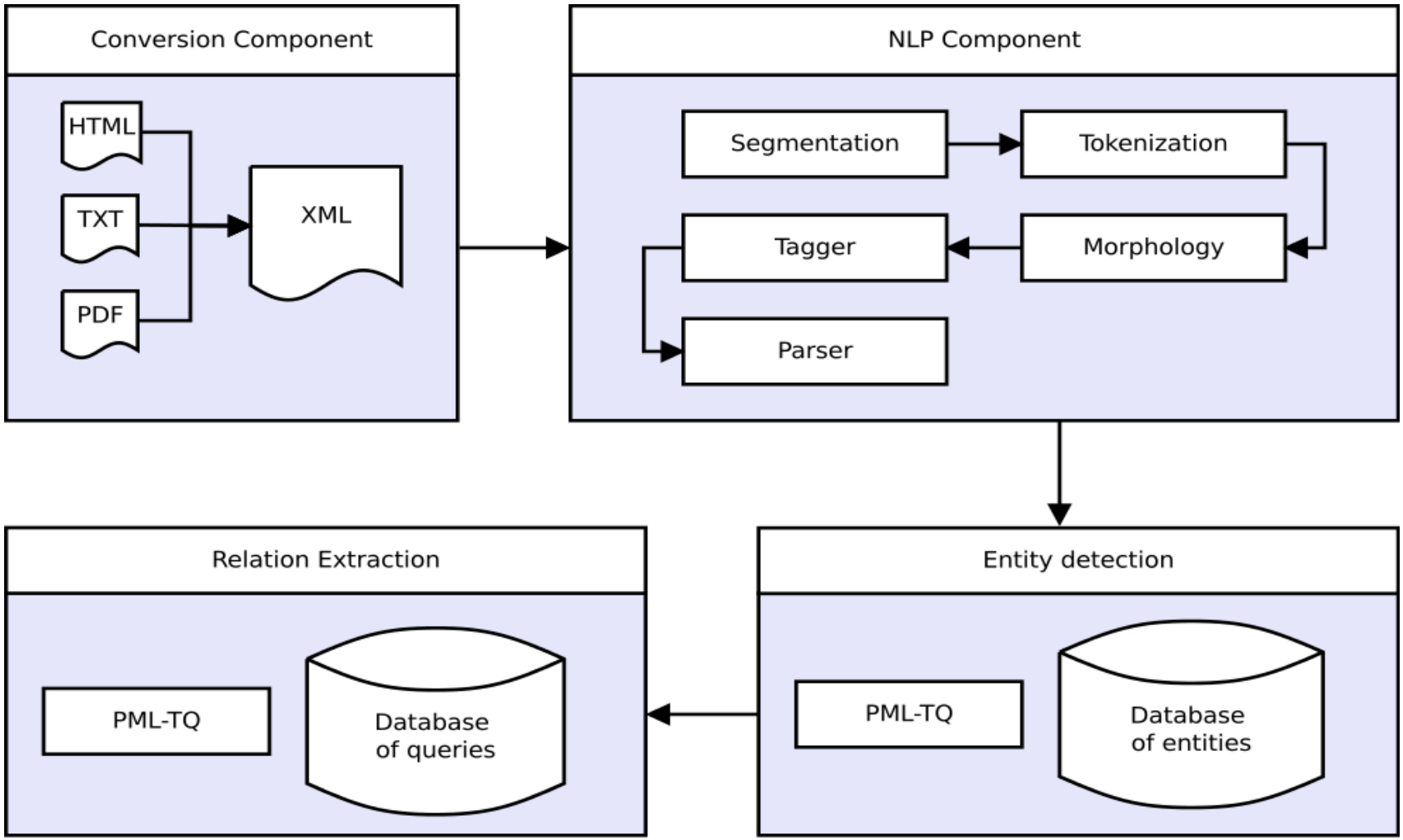
Sentences

- simple sentences are very rare
- usually long and very complex

Legal texts are “generally considered very difficult to read and understand”. (Tiersma, 2010)

RExtractor Architecture

Adaptation for legislative domain

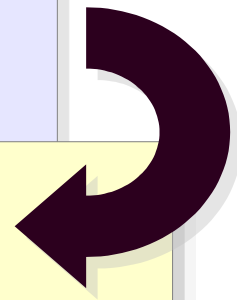


Conversion component

```
HLAVA I
ÚVODNÍ USTANOVENÍ

§ 1
Předmět úpravy

Tato vyhláška zapracovává příslušné předpisy Evropské unie a upravuje:
   a) způsob vymezení hydrogeologických rajonů, vymezení útvarů podzemních vod,
   b) způsob hodnocení stavu podzemních vod a
   c) náležitosti programů zjišťování a hodnocení stavu podzemních vod.
```

Conversion component

```
HLAVA I
ÚVODNÍ USTANOVENÍ

§ 1
Předmět úpravy

Tato
   a
   b
   c
```



```
<head id="11" label="HLAVA I">
        <title>ÚVODNÍ USTANOVENÍ</title>
        <section id="12" label="§ 1">
                <title>Předmět úpravy</title>
                <text>Tato vyhláška zapracovává příslušné
                      předpisy Evropské unie a upravuje:</text>
                <section id="13" label="a)">
                        <text>způsob vymezení hydrogeologických rajonů,
                              vymezení útvarů podzemních vod,</text>
                </section>
                <section id="14" label="b)">
                        <text>způsob hodnocení stavu podzemních vod a</text>
                </section>
                <section id="15" label="c)">
                        <text>náležitosti programů zjišťování a
                              hodnocení stavu podzemních vod.</text>
                </section>
        </section>
</head>
```

NLP Component

Corpus of Czech legal texts (CCLT)

- Accounting Act (563/1991 Coll.)
- Decree on Double-entry Accounting for undertakers (500/2002 Coll.)
- automatically parsed, then manually checked
  - 1,133 manually annotated a-trees
  - 35,085 tokens
  - Credit to Zdeňka Urešová

NLP Component

Corpus of Czech legal texts (CCLT)

- enumerations and lists as one tree
- manual annotation guidelines
  - split sentences according to formal markers
  - use links for dependencies between partial trees
- automatic procedure merges partial annotations into a final tree

Pipeline visualization available on-line at
ufal.mff.cuni.cz/intlib

NLP Component

Automatic parsers for Czech

- trained on newspaper texts
- verification whether we can use the parser trained on newspaper texts or some modifications are needed
- MST parser Ryan McDonald, Fernando Pereira, Kiril Ribarov, Jan Hajič (2005): Non-projective Dependency Parsing using Spanning Tree Algorithms. In: Proceedings of HLT/EMNLP, Vancouver, British Columbia.

NLP Component

Sentence splitting

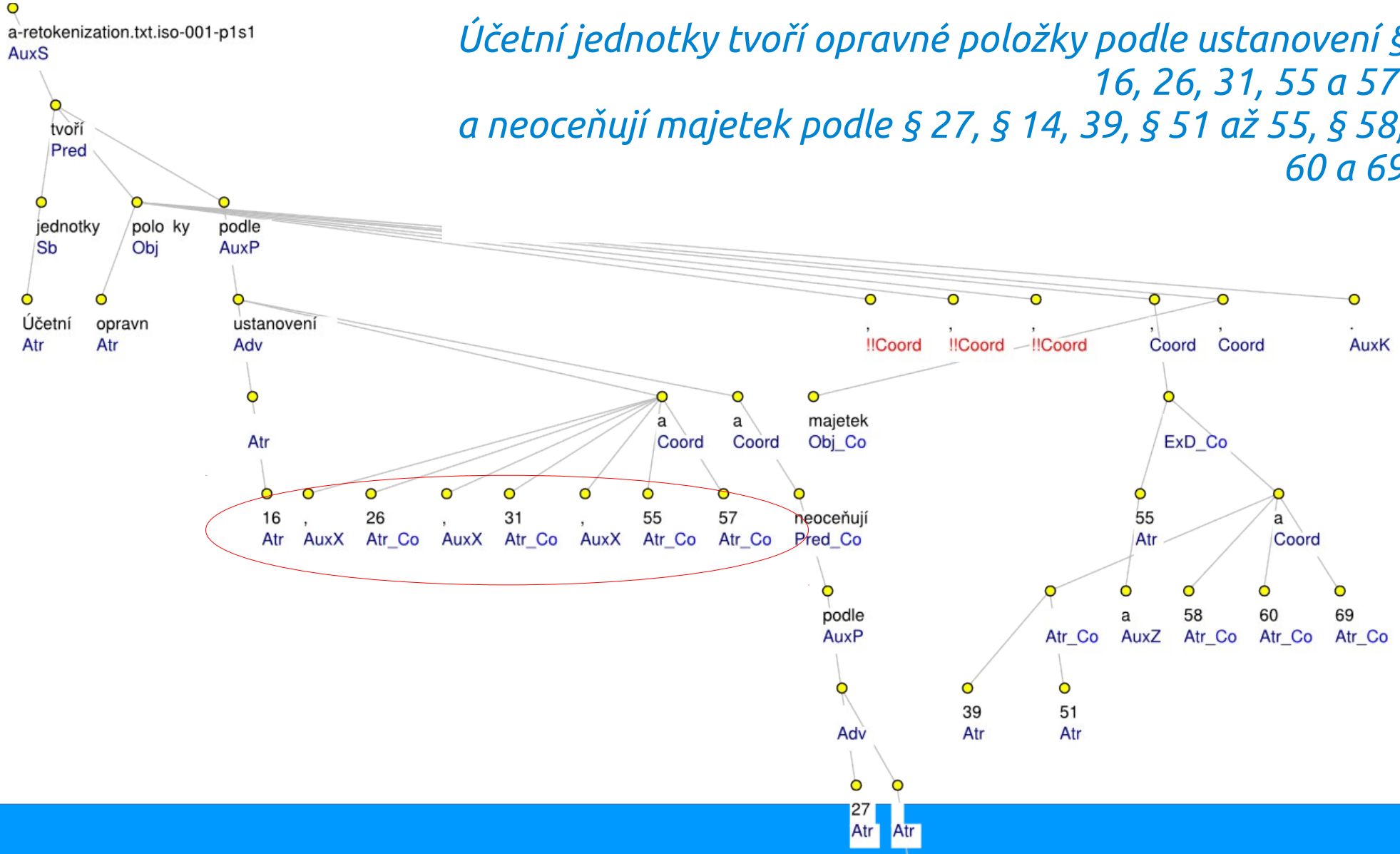
- We substitute long lists and enumerations by several shorter sentences

Original sentence	New sentences
(2) Veřejným rozpočtem se pro účely tohoto zákona rozumí a) státní rozpočet b) rozpočet státního fondu, c) rozpočet Evropské unie, nebo d) rozpočet, o němž to stanoví zákon.	Veřejným rozpočtem se pro účely tohoto zákona rozumí státní rozpočet.
	Veřejným rozpočtem se pro účely tohoto zákona rozumí rozpočet státního fondu.
	Veřejným rozpočtem se pro účely tohoto zákona rozumí rozpočet Evropské unie.
	Veřejným rozpočtem se pro účely tohoto zákona rozumí rozpočet, o němž to stanoví zákon.

NLP Component

Re-tokenization

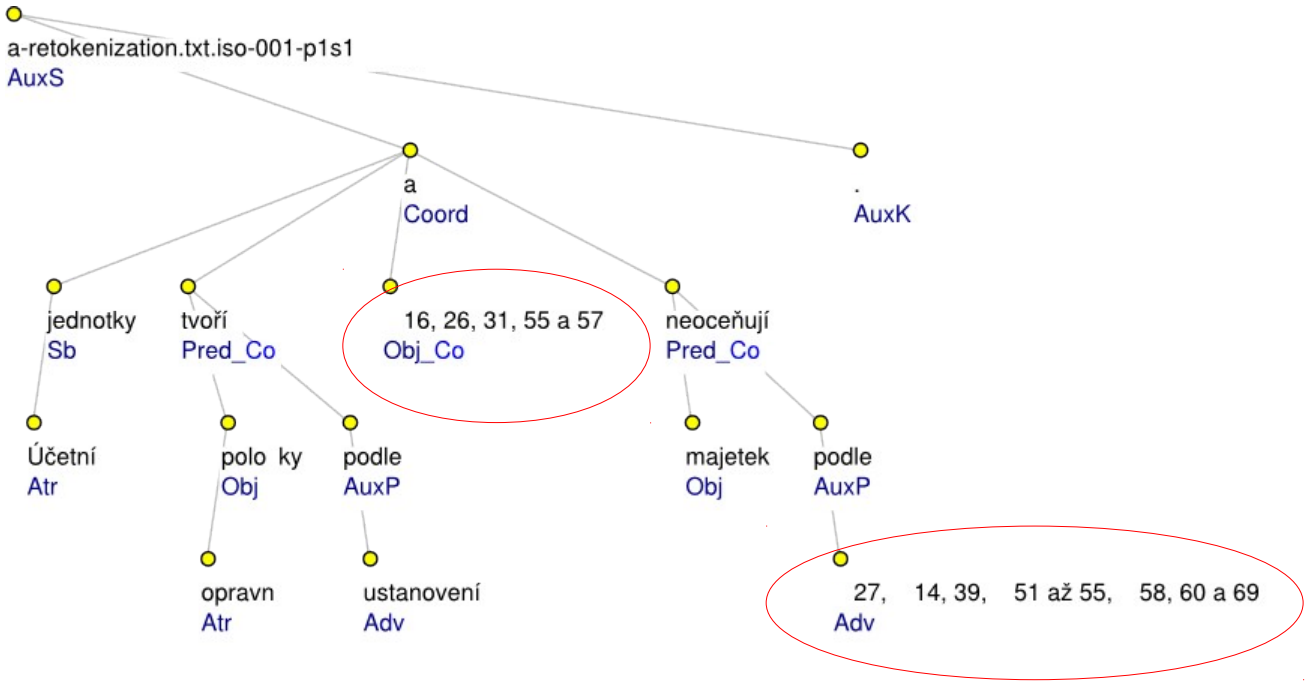
Účetní jednotky tvoří opravné položky podle ustanovení § 16, 26, 31, 55 a 57 a neoceňují majetek podle § 27, § 14, 39, § 51 až 55, § 58, 60 a 69



NLP Component

Re-tokenization

Účetní jednotky tvoří opravné položky podle ustanovení ***§ 16, 26, 31, 55 a 57***
a neoceňují majetek podle ***§ 27, § 14, 39, § 51 až 55, § 58, 60 a 69***



Entity Detection Component

Entities in CCLT

- Accounting subdomain
- Entities manually annotated by Sysnet, Ltd.
  - Decree on Double-entry Accounting for undertakers (500/2002 Coll.)

Sample

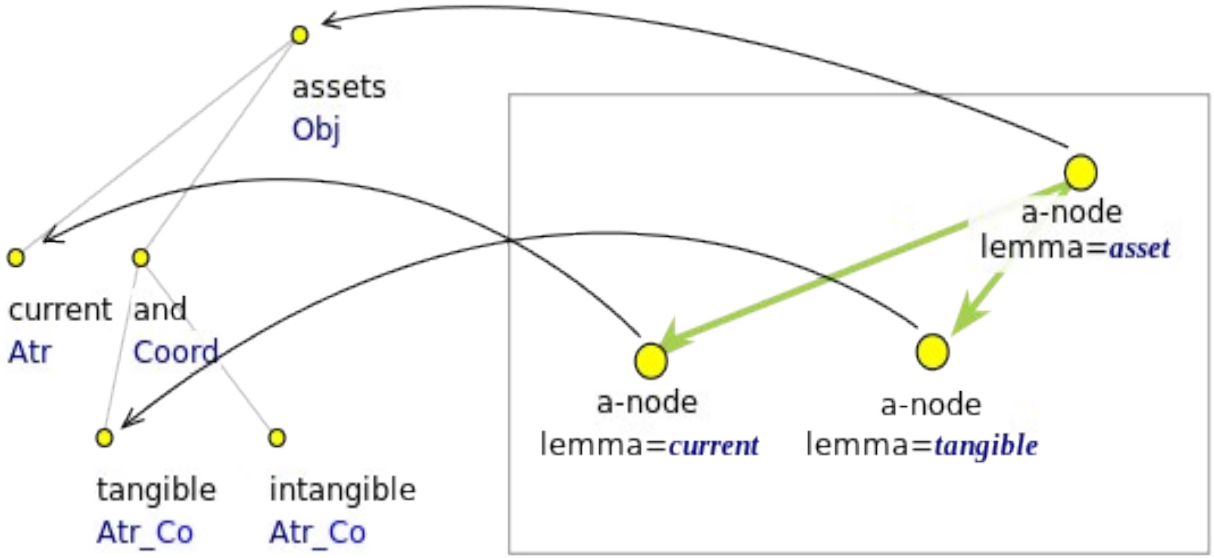
___(1) Vyhláška se vztahuje na účetní jednotky podle § 1 odst. 2 písm. a) a b) zákona, s výjimkou účetních jednotek uvedených v odstavci 2, a na účetní jednotky podle § 1 odst. 2 písm. d) až h) zákona.

___(2) Z účetních jednotek uvedených v odstavci 1 se tato vyhláška nevztahuje na účetní jednotky podle § 19a zákona, pokud zvláštní právní předpis 1c) nestanoví jinak, a na účetní jednotky, jejichž účetnictví upravuje zvláštní právní předpis 1d). Dále se tato vyhláška, s výjimkou § 62 odst. 2 až 5, nevztahuje na účetní jednotky podle § 23a zákona.

Entity Detection Component

Initializing DBE with entities from CCLT

- Each (unique) entity parsed automatically by MST
- Automatic procedure takes an entity dependency tree and creates PML-TQ query



Entity Detection Component

Experiment

- identify entities in the gold standard trees in CCLT
  - with re-tokenized tokens and *(very) long* sentences
- identify entities in the trees created by MST
  - with re-tokenized tokens and split sentences

Parsing method	Extracted	TP	FP	FN	Precision	Recall
Manual	16428	9549	6879	628	58.1	93.8
Automatic	16160	9278	6882	838	57.4	91.7

Results

- high False positives
- automatic parser has low influence on detection

Relation Extraction Component

Types of relations

- Definitions (D) – entities are defined or explained
  - *Náhradním ubytováním se rozumí byt o jedné místnosti nebo pokoj ve svobodárně nebo podnájem v zařízené nebo nezařízené části bytu jiného nájemce.*
- Obligations (O) – an entity is obligated to do something
  - *K návrhu je navrhovatel povinen připojit listiny , kterých se v návrhu dovolává.*
- Rights (R) – an entity has right to do something
  - *Nabyvatel může uplatňovat nárok z odpovědnosti za vady u soudu jen tehdy , vytkl-li vady bez zbytečného odkladu po té , kdy měl možnost věc prohlédnout .*

Relation Extraction Component

Manual design of queries

- Strategy: cover maximum of relations with minimum of queries
- tree query expert
  - observes typical constructions for a given relation type
  - designs a query for the most frequent construction
  - goes through matches and redesigns the query if needed

Relation Extraction Component

Query design & evaluation on CCLT

- Query design
  - on Accounting Act (563/1991 Coll.)
  - 5 queries for Definitions
  - 4 queries for Rights
  - 2 queries for Obligation
- Evaluation
  - on Decree on Double-entry Accounting for undertakers (500/2002 Coll.)

Relation Extraction Component

Results

	D	R	O	Total
# of queries	5	4	2	11
Goldstandard	97	308	62	467
Extracted	70	255	41	366
True positive	53	206	36	295
False negative	44	102	26	172
False positive	17	49	5	71
Precision (%)	**75.7**	**80.8**	**87.8**	**80.6**
Recall (%)	**54.6**	**66.9**	**58.1**	**63.2**

Relation Extraction Component

Error analysis

Error	# of errors	Ratio
Parser	145	59.7%
Query	93	38.3%
Entity	5	2.1%

Results

- errors in automatic parsing
- query design

Relation Extraction Component

Experiment with more data

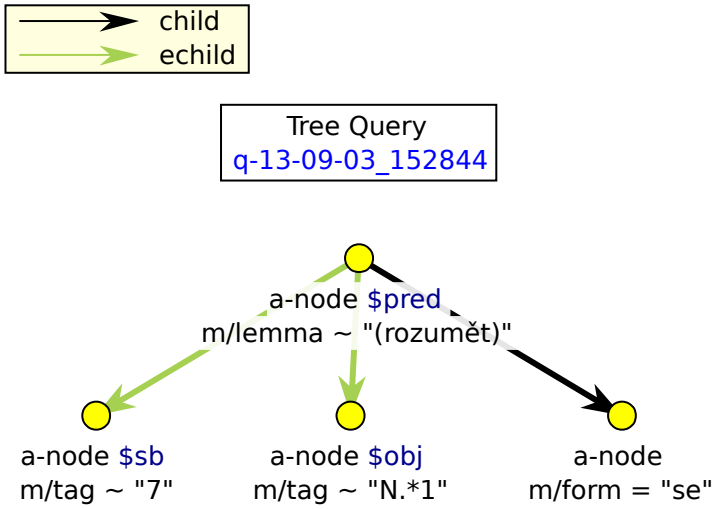
- 28 laws from the accounting subdomain
- 27,808 sentences
- 745,137 tokens

D		R		O	
D_1	36	R_1	240	O_1	183
D_2	287	R_2	470	O_2	37
D_3	35	R_3	127		
D_4	466	R_4	6		
D_5	46				
Total	1580	Total	843	Total	220

Relation Extraction Component

Query example - Definition

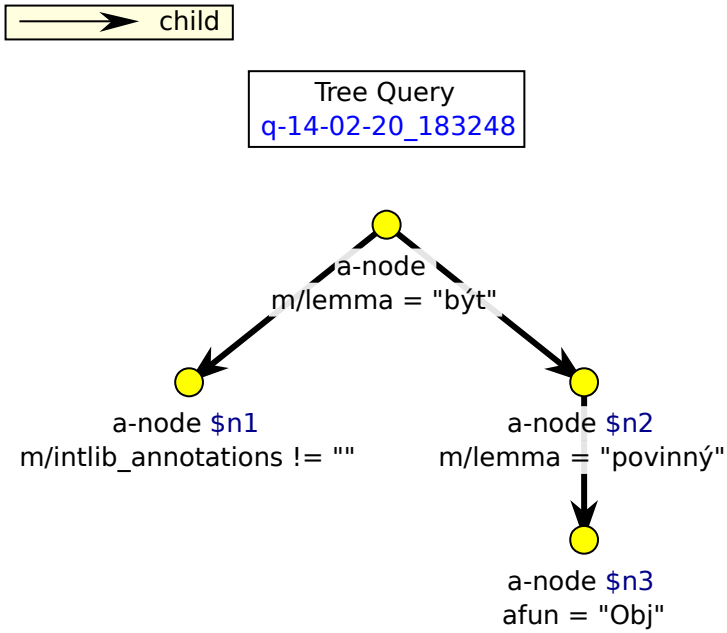
– *Náhradním ubytováním se rozumí byt o jedné místnosti nebo pokoj ve svobodárně nebo podnájem v zařízené nebo nezařízené části bytu jiného nájemce .*



Relation Extraction Component

Query example – Obligation

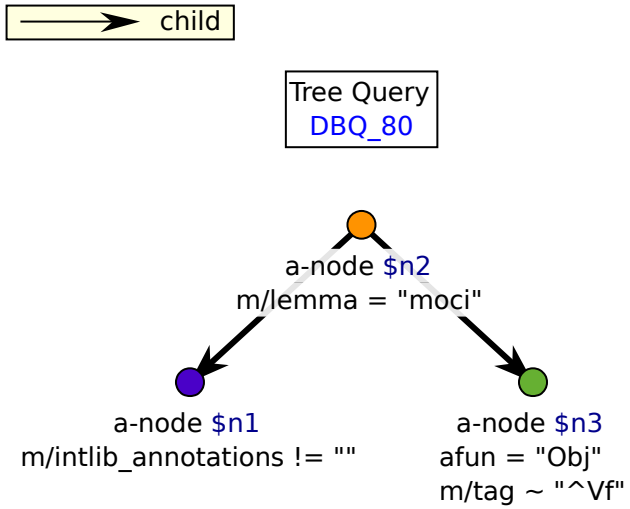
- *K návrhu je navrhovatel povinen připojit listiny , kterých se v návrhu dovolává .*



Relation Extraction Component

Query example – Right

- *Nabyvatel může uplatňovat nárok z odpovědnosti za vady u soudu jen tehdy , vytkl-li vady bez zbytečného odkladu po té , kdy měl možnost věc prohlédnout .*



Future Work

Legislative domain

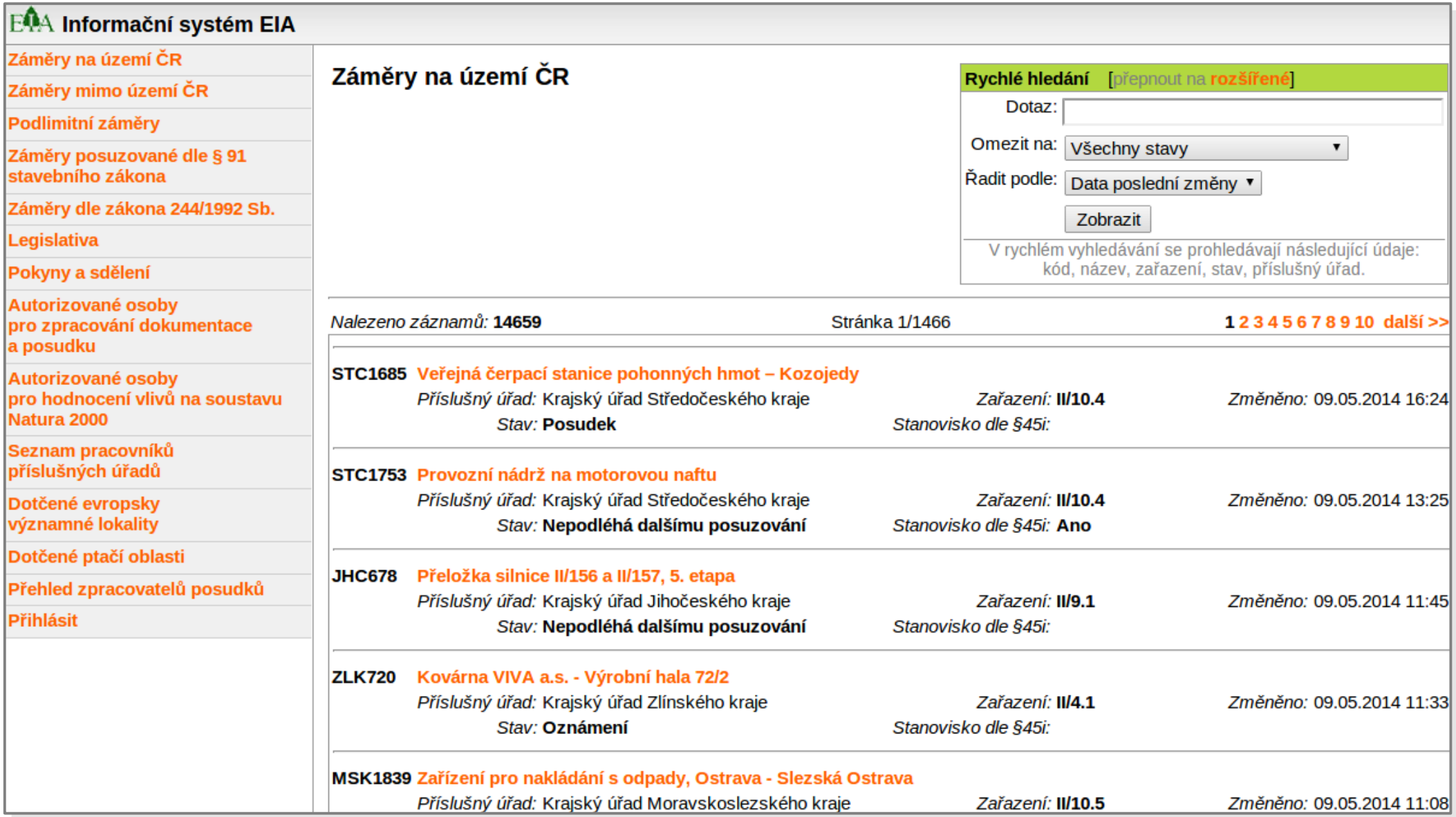
- Parsing
  - evaluation and adaptation
- Entity detection
  - automatic entity detection based on a sample of entities annotated manually
- Relation extraction
  - automatic query design

Case study on environmental domain

Case study on environmental domain

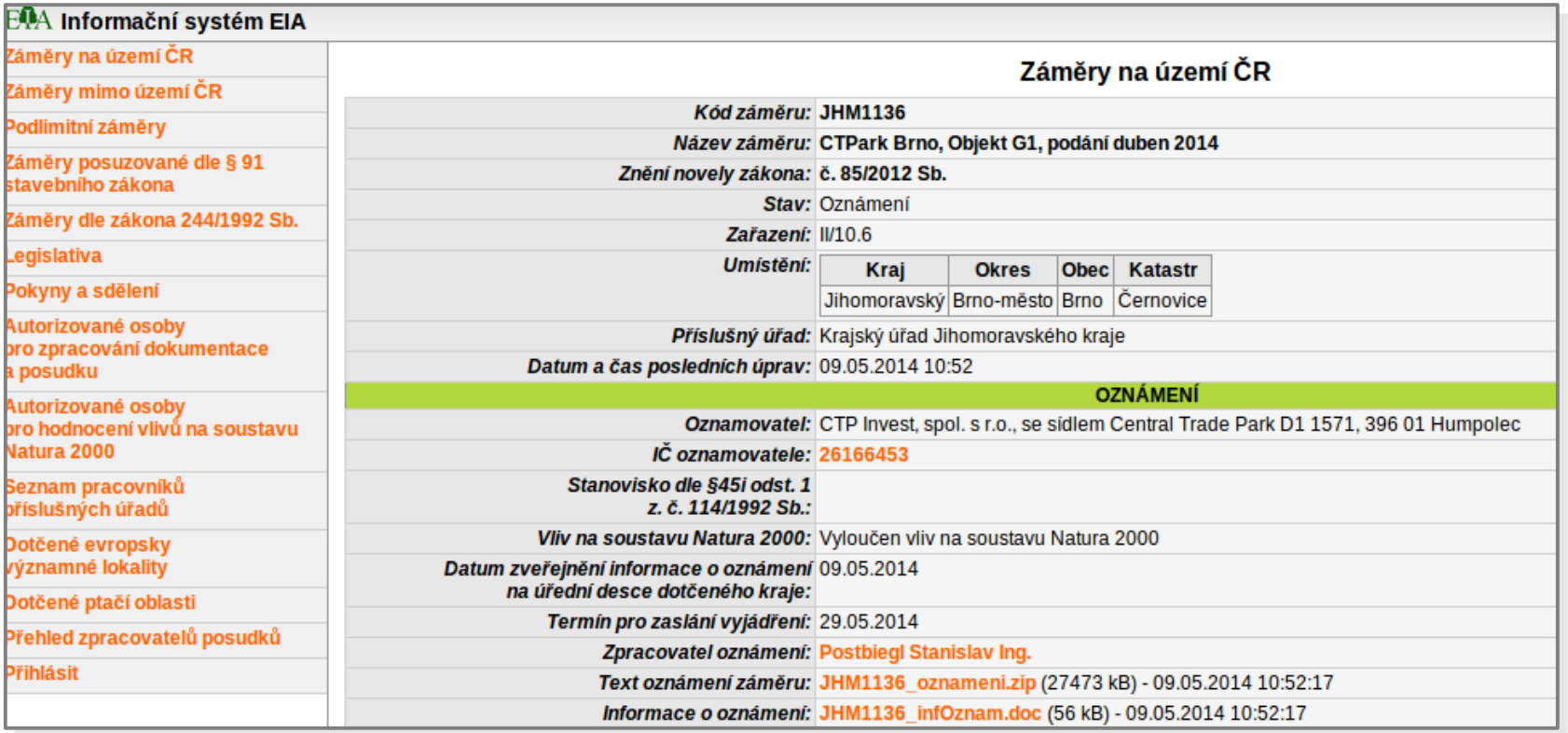
- What are the environmental consequences of a project?
- Environmental Impact Assessment considers the environmental impacts whether or not to proceed with a project.
- In the Czech Republic, CENIA administers the EIA information system.

EIA system


Informační systém EIA
Záměry na území ČR
Záměry mimo území ČR
Podlimitní záměry
Záměry posuzované dle § 91 stavebního zákona
Záměry dle zákona 244/1992 Sb.
Legislativa
Pokyny a sdělení
Autorizované osoby pro zpracování dokumentace a posudku
Autorizované osoby pro hodnocení vlivů na soustavu Natura 2000
Seznam pracovníků příslušných úřadů
Dotčené evropsky významné lokality
Dotčené ptačí oblasti
Přehled zpracovatelů posudků
Přihlásit
Záměry na území ČR
Rychlé hledání [přepnout na rozšířené]
Dotaz:
Omezit na: Všechny stavy
Řadit podle: Data poslední změny
Zobrazit
V rychlém vyhledávání se prohledávají následující údaje: kód, název, zařazení, stav, příslušný úřad.
Nalezeno záznamů: 14659
Stránka 1/1466
1 2 3 4 5 6 7 8 9 10 další >>
STC1685 Veřejná čerpací stanice pohonných hmot – Kozojedy
Příslušný úřad: Krajský úřad Středočeského kraje
Zařazení: II/10.4
Změněno: 09.05.2014 16:24
Stav: Posudek
Stanovisko dle §45i:
STC1753 Provozní nádrž na motorovou naftu
Příslušný úřad: Krajský úřad Středočeského kraje
Zařazení: II/10.4
Změněno: 09.05.2014 13:25
Stav: Nepodléhá dalšímu posuzování
Stanovisko dle §45i: Ano
JHC678 Přeložka silnice II/156 a II/157, 5. etapa
Příslušný úřad: Krajský úřad Jihočeského kraje
Zařazení: II/9.1
Změněno: 09.05.2014 11:45
Stav: Nepodléhá dalšímu posuzování
Stanovisko dle §45i:
ZLK720 Kovárna VIVA a.s. - Výrobní hala 72/2
Příslušný úřad: Krajský úřad Zlínského kraje
Zařazení: II/4.1
Změněno: 09.05.2014 11:33
Stav: Oznámení
Stanovisko dle §45i:
MSK1839 Zařízení pro nakládání s odpady, Ostrava - Slezská Ostrava
Příslušný úřad: Krajský úřad Moravskoslezského kraje
Zařazení: II/10.5
Změněno: 09.05.2014 11:08

Example

- Amazon's plan to build a distribution center in Brno, CR (no, no, no, yes by Brno councilors)
- May 9, 2014: the new intention posted at EIA by CTP Invest



Informační systém EIA

- Záměry na území ČR
- Záměry mimo území ČR
- Podlimitní záměry
- Záměry posuzované dle § 91 stavebního zákona
- Záměry dle zákona 244/1992 Sb.
- Legislativa
- Pokyny a sdělení
- Autorizované osoby pro zpracování dokumentace a posudku
- Autorizované osoby pro hodnocení vlivů na soustavu Natura 2000
- Seznam pracovníků příslušných úřadů
- Dotčené evropsky významné lokality
- Dotčené ptačí oblasti
- Přehled zpracovatelů posudků
- Přihlásit

Záměry na území ČR

Kód záměru:	**JHM1136**
Název záměru:	**CTPark Brno, Objekt G1, podání duben 2014**
Znění novely zákona:	**č. 85/2012 Sb.**
Stav:	Oznámení
Zařazení:	II/10.6
Umístění:	Kraj: Jihomoravský; Okres: Brno-město; Obec: Brno; Katastr: Černovice
Příslušný úřad:	Krajský úřad Jihomoravského kraje
Datum a čas posledních úprav:	09.05.2014 10:52
OZNÁMENÍ	
Oznamovatel:	CTP Invest, spol. s r.o., se sídlem Central Trade Park D1 1571, 396 01 Humpolec
IČ oznamovatele:	26166453
Stanovisko dle §45i odst. 1 z. č. 114/1992 Sb.:	
Vliv na soustavu Natura 2000:	Vyloučen vliv na soustavu Natura 2000
Datum zveřejnění informace o oznámení na úřední desce dotčeného kraje:	09.05.2014
Termín pro zaslání vyjádření:	29.05.2014
Zpracovatel oznámení:	Postbiegl Stanislav Ing.
Text oznámení záměru:	JHM1136_oznameni.zip (27473 kB) - 09.05.2014 10:52:17
Informace o oznámení:	JHM1136_infOznam.doc (56 kB) - 09.05.2014 10:52:17

Mining EIA documentation

- Sysnet, Ltd. specified what entities and relations to extract, e.g.
  - Title (Section B.I.1)
  - Category, type (Section B.I.1)
  - Capacity, size (Section B.I.2, B.I.6)
  - Location (Section B.I.3)
  - Scheduling (Section B.I.7)
  - ...

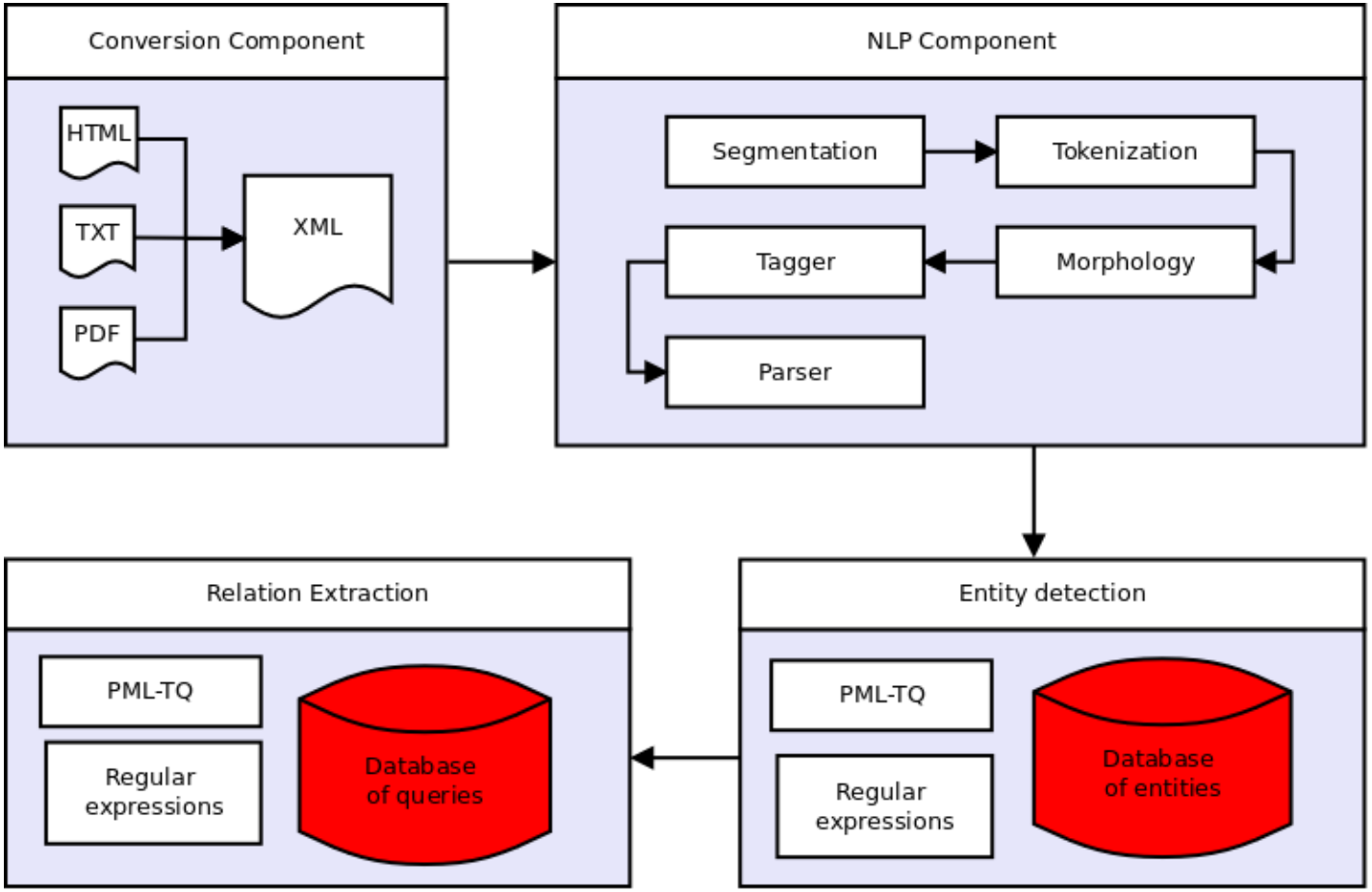
Focus on section B.I.2

- **Example**

  Vlastní areál bude sestávat z halového objektu o ploše cca 96 000 m² , který bude uvnitř rozdělen na 3 haly … Předpokládají se 2 krytá stání pro jízdní kola a 1150 parkovacích stání pro osobní vozidla … Součástí záměru je realizace sadových úprav, která zahrnuje výsadbu více než 250 ks vzrostlých stromů

  - *The park will consists of a hall with the area of cca 96 000 m² that will be split into 3 halls … There will be 2 roofed bicycle parking stations and 1,150 parking slots …*

Using RExtractor



- queries by regular expressions

Regular expressions

Dále je provozována produkční stáj VKK pro 336 ks dojnic (403,2 DJ). (*In addition, a reproductive barn VKK is used for 336 cows.*)

(Adj Nom)? (Noun Nom) (number) (unit) (Noun Gen)

(attribute)(entity) (number) (unit) (entity)

(*reproductive*)(*barn*) (*336*) (*pcs*) (*cow*)

Credit to Ivana Lukšová

Both l. & e. domain

- Evaluation
  - Developers vs. users
  - Gold standard data vs. practical use cases
  - Experience vs. expectation
  - Scientific contribution vs. “making life easier”