Extrakce informací pomocí GATE

Malostranská IT setkání

29.5.2012

Jan Dědek

GATE

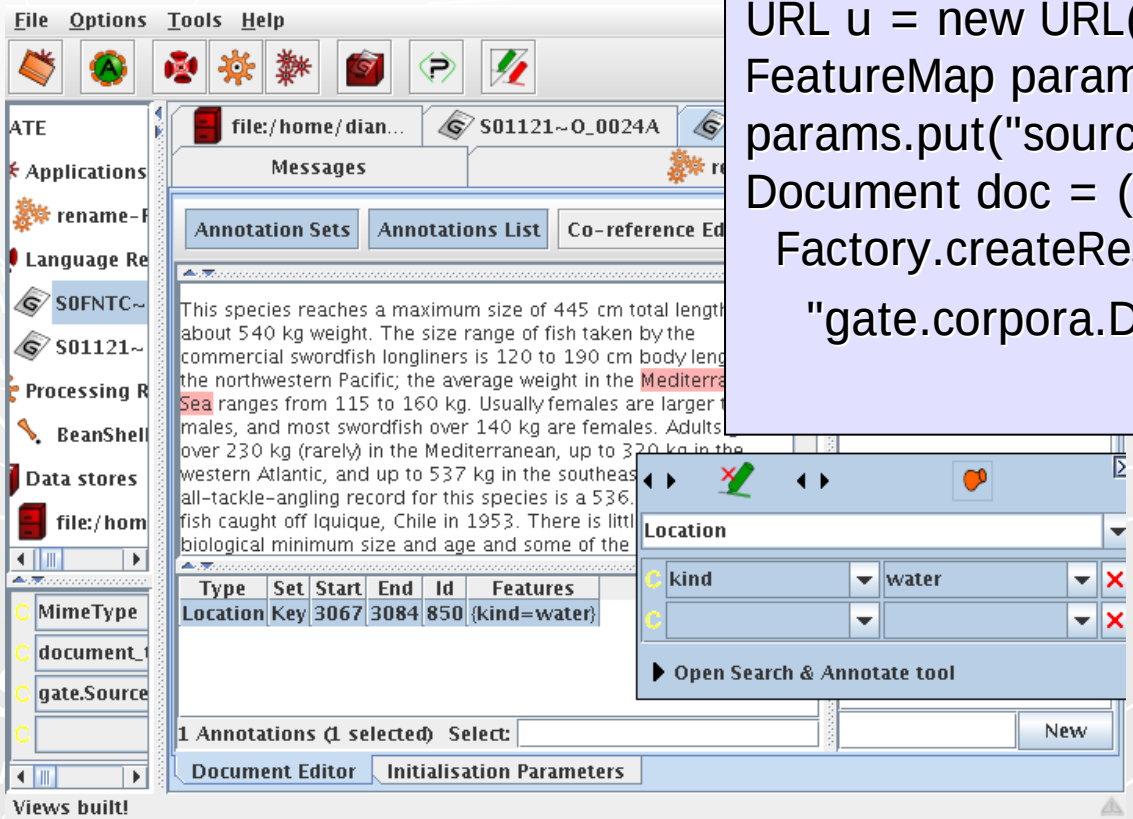
general architecture for text engineering

- The University of Sheffield
- Java
- Open Source
- http://gate.ac.uk/

The University Of Sheffield.

IDE nebo API

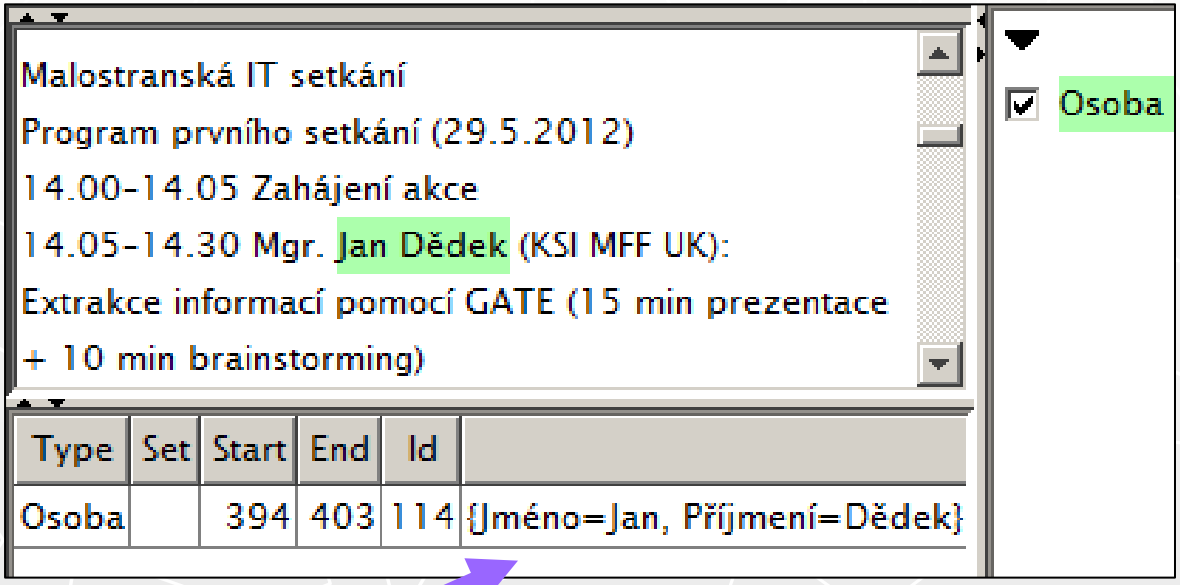
- GATE Developer
- GATE Embedded




```
URL u = new URL("http://www.fsolutions.cz/");
FeatureMap params = Factory.newFeatureMap();
params.put("sourceUrl", u);
Document doc = (Document)
  Factory.createResource(
    "gate.corpora.DocumentImpl", params);
```

Anotace v GATE - tagování

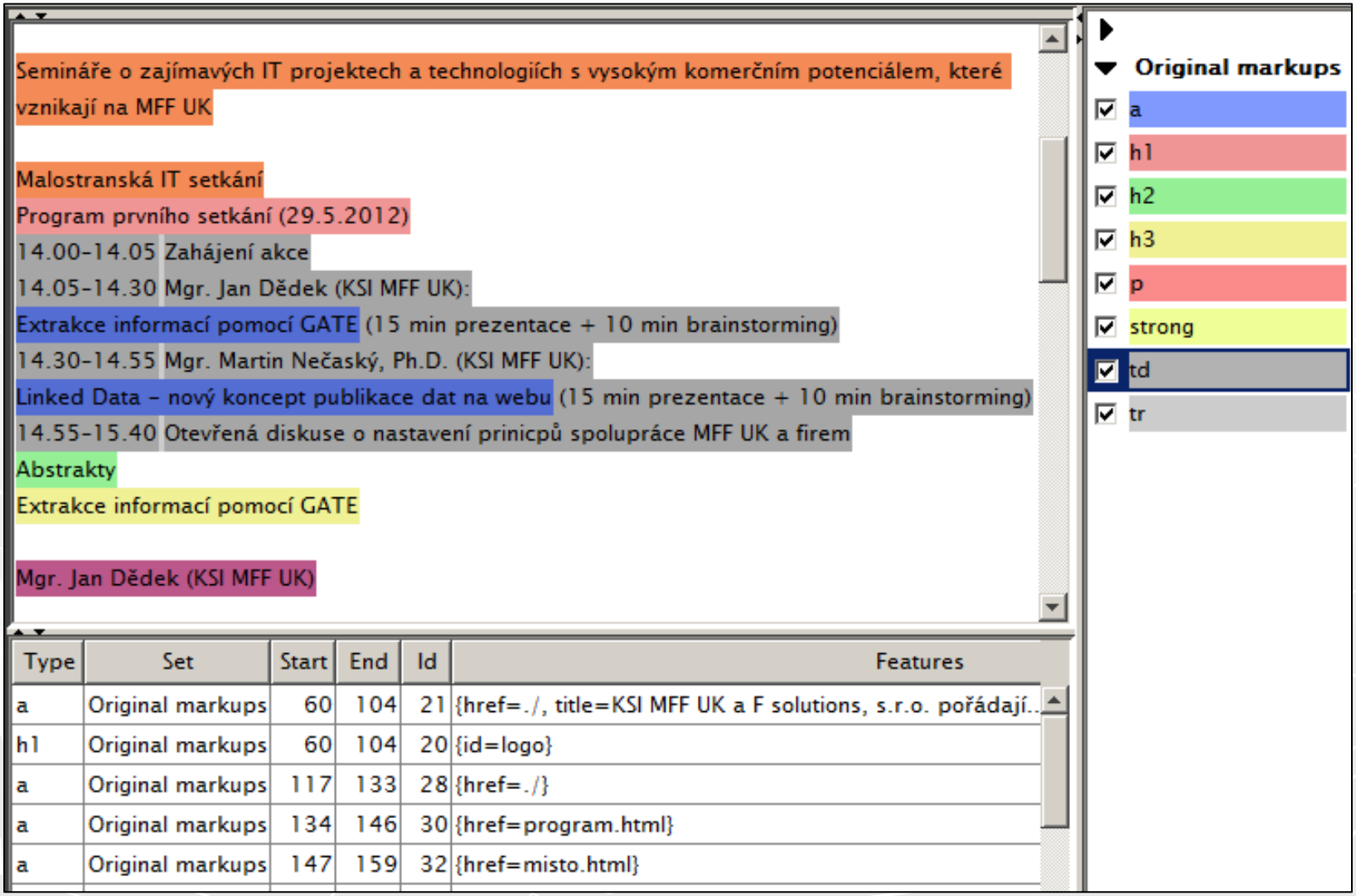
- Základ veškeré práce v GATE
- Anotace
  - Kousek textu
  - Id
  - Začátek
  - Konec
  - Typ
  - Vlastnosti (Features)
  - A to je celé!



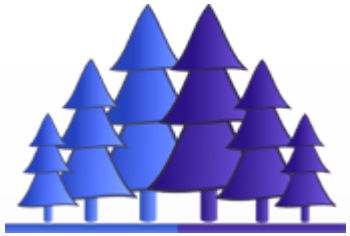
Jak se tam anotace dostanou?

- Ručně (annotation editor)
- Z importu (značky html, xml, pdf - málo, ...)
- Dalším zpracováním
  - Segmentace
  - Tokenizace
  - Morfologická analýza (lemmatizace/stemming)
  - Syntaktická analýza (parsing)
  - Gazetteer (výčty)
  - JAPE – regulární výrazy nad anotacemi
  - Strojové učení

Import HTML



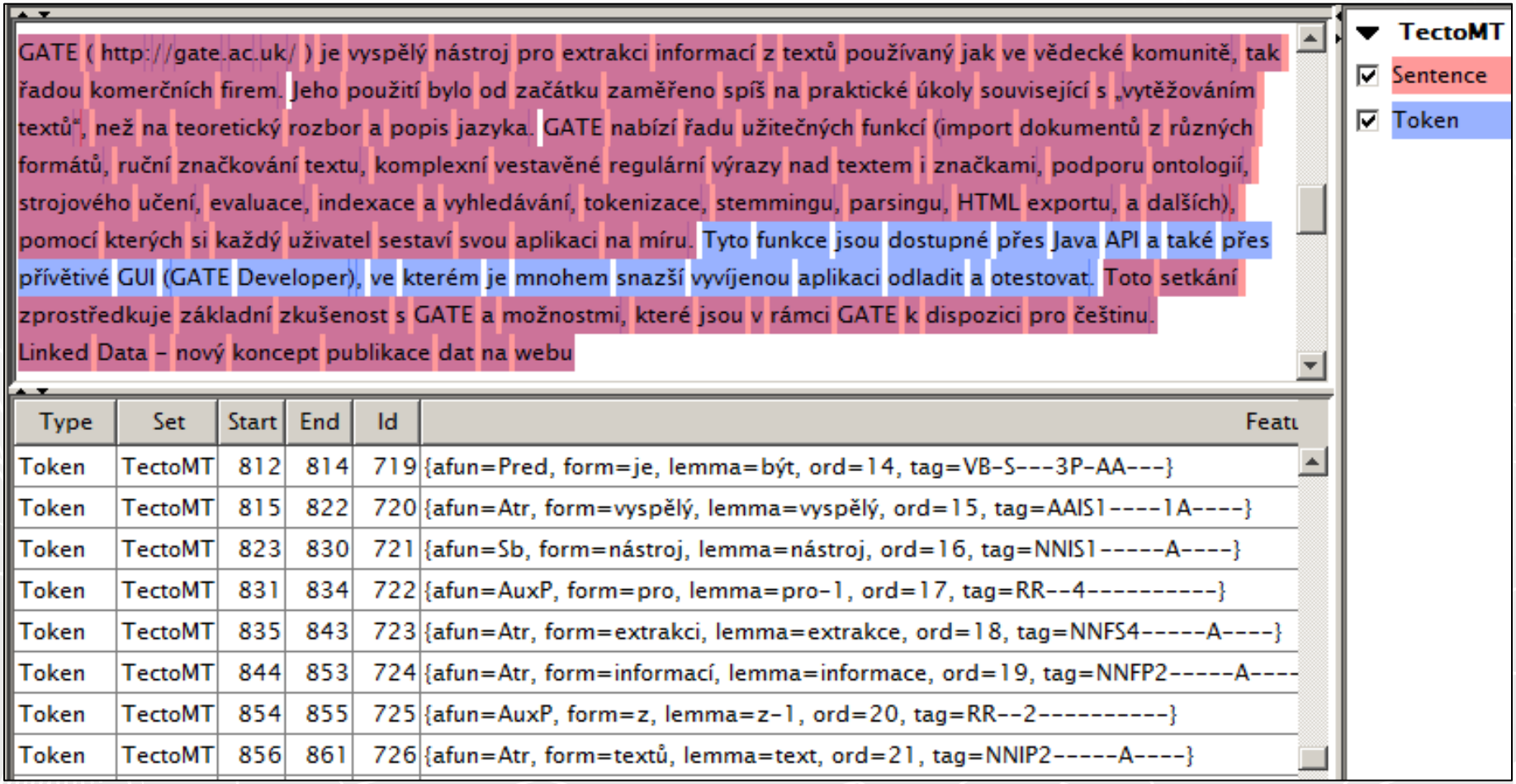
Type	Set	Start	End	Id	Features
a	Original markups	60	104	21	{href=./, title=KSI MFF UK a F solutions, s.r.o. pořádají..
h1	Original markups	60	104	20	{id=logo}
a	Original markups	117	133	28	{href=./}
a	Original markups	134	146	30	{href=program.html}
a	Original markups	147	159	32	{href=misto.html}



Čeština v GATE

- TectoMT (Treex)
  - Poskytuje jednotný (Perl) interface ke spoustě lingvistických nástrojů nejen pro Češtinu.
  - http://ufal.mff.cuni.cz/tectomt/
  - http://ufal.mff.cuni.cz/treex/
- Czsem Mining Suite
  - Umožňuje používat TectoMT uvnitř GATE.
  - ... mimo jiné :-)
  - Jan Dědek a kol.
  - http://czsem.berlios.de/

Segmentace, Tokenizace



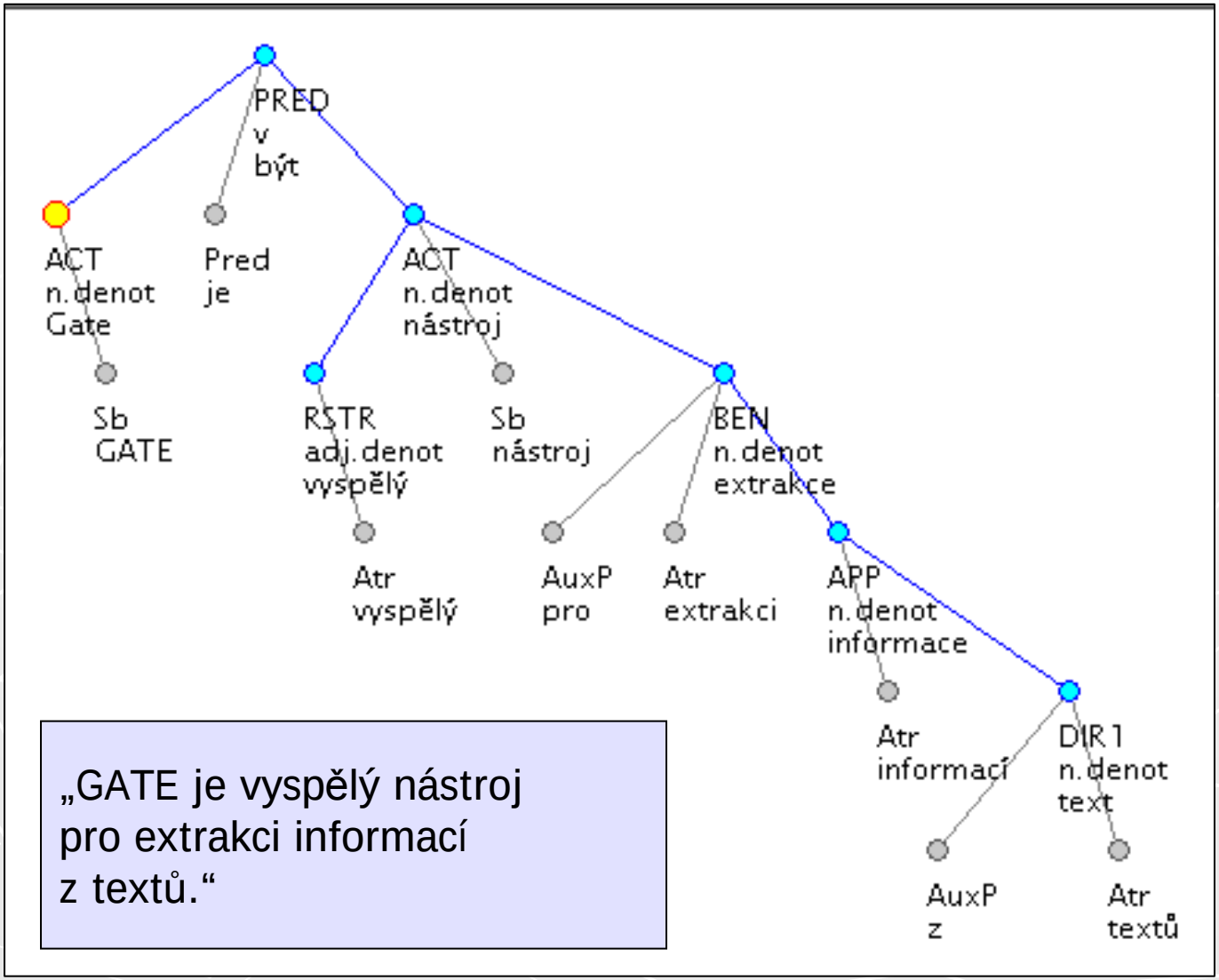
Morfologická analýza

form=GATE	lemma=Gat_;G	tag=NNIS5-----A----
form=je	lemma=být	tag=VB-S---3P-AA---
form=vyspělý	lemma=vyspělý	tag=AAIS1----1A----
form=nástroj	lemma=nástroj	tag=NNIS1-----A----
form=pro	lemma=pro-1	tag=RR--4----------
form=extrakci	lemma=extrakce	tag=NNFS4-----A----
form=informací	lemma=informace	tag=NNFP2-----A----
form=z	lemma=z-1	tag=RR--2----------
form=textů	lemma=text	tag=NNIP2-----A----

- Lemma
- Morfologická značka
  - A - Adjective
  - C - Numeral
  - D - Adverb
  - I - Interjection
  - J - Conjunction
  - N - Noun
  - P - Pronoun
  - V - Verb
  - R - Preposition
  - T - Particle
  - X - Unknown, Not Determined
  - **Atd...**
    - Osoba, číslo, čas, rod, pád ...

http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/m-layer/html/index.html

Syntaktická analýza (parsing)



„GATE je vyspělý nástroj
pro extrakci informací
z textů.“

Gazetteer (výčty) (1)

- Pozor měníme „running example“...
- Mějme databázi **lékařských článků**
- a **slovník lékařských pojmů**

- GATE Gazetteer anotuje všechny výskyty lékařských pojmů ve článku
  - Včetně skloňování, díky předchozí morfologické analýze

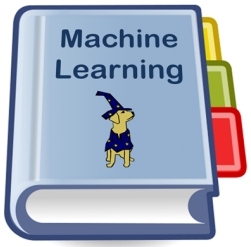
Gazetteer (výčty) (2)

Technika vyšetření artrografie se provádí u pacienta ležícího na zádech, zásadně za skiaskopické kontroly. Paže je natažená v supinaci. V lokální anestezii, dlouhou tenkou jehlou, používanou k lumbální punkci, s napojenou hadičkou vpichujeme do oblasti dolního vnitřního kvadrantu hlavice (ne tedy přímo do kloubní štěrbiny – v tom případě bychom narazili na labrum) kontrastní látku, obvykle 8–10 ml a doplníme 10 ml vzduchu. Rameno lehce rozcvičíme, aby se kontrast dostal do všech záhybů kloubní dutiny. Nato rameno osnímkujeme přehledně v pronaci, supinaci, abdukci a addukci. Snímek vstoje, kdy se aplikovaný vzduch hromadí v horních oddílech kloubu, usnadní průkaz ruptury rotátorové manžety. Následuje CT vyšetření 3mm scany se zvětšeným obrazem, v kostním okně, v neutrální poloze.

Type	Set	Start	End	Id	Features
Lookup	llong	1669	1680	1709	{czTerm=artrografie, enTerm=Arthrography, majorType=czmesh_lemmas, mesh
Lookup	llong	1694	1702	1710	{czTerm=pacienti, enTerm=Patients, majorType=czmesh_lemmas, meshID=D01
Lookup	llong	1715	1721	1711	{czTerm=záda, enTerm=Back, majorType=czmesh_lemmas, meshID=D001415,
Lookup	llong	1757	1761	1712	{czTerm=paže, enTerm=Arm, majorType=czmesh_lemmas, meshID=D001132,
Lookup	llong	1776	1784	1713	{czTerm=supinace, enTerm=Supination, majorType=czmesh_lemmas, meshID=D
Lookup	llong	1822	1828	1714	{czTerm=jehly, enTerm=Needles, majorType=czmesh_lemmas, meshID=D0093
Lookup	llong	1852	1858	1715	{czTerm=punkce, enTerm=Punctures, majorType=czmesh_lemmas, meshID=D0
Lookup	llong	2017	2033	1716	{czTerm=kontrastní látky, enTerm=Contrast Media, majorType=czmesh_lemmas,

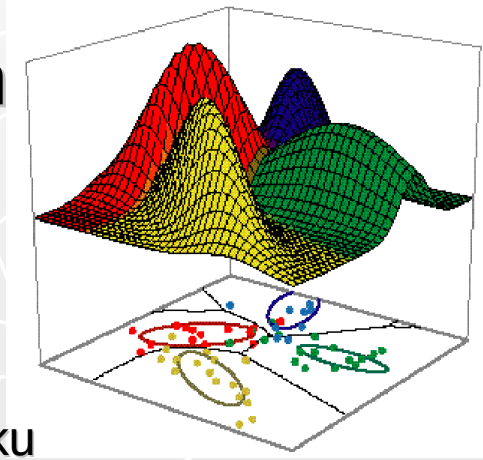
JAPE (Java Annotation Patterns Engine) regulární výrazy nad anotacemi

- K anotacím *příjmení* přidej anotaci *vysokoškolák* pokud jim předchází *titul*.
- Anotuj všecky IP adresy začínající „192.168“ jako *domácí*.
- ... všechny *emailové adresy* končící „cuni.cz“ jako *pracovní.*
- *Věty* obsahující anotaci *chyba* označ jako *důležité.*



Strojové učení

- Nutné vytvořit trénovací sadu
  - Obvykle ručně označkovat větší množství dokumentů
    - Kolik? Záleží na variabilitě extrahované informace.
    - Každá forma výskytu musí být pokryta nějakým trénovacím příkladem.
- Naučený algoritmus funguje na nových dokumentech
  - Doplňuje nové anotace nebo
  - Klasifikuje existující
    - Možné použít pro „sentiment analysis“
      - Analýza spokojenosti, emočního náboje příspěvku
  - Vhodné provést evaluaci na zbytku ručních anotací
    - Velmi snadné v GATE



Co s tím?
K čemu to všecko je?

- Je to na Vás!
  - Vytvořte si (nechte si udělat) analýzu v GATE Vašich dokumentů, logů, webových komentářů, emailů, ...
  - A naložte s jejím výstupem jak potřebujete.
- Indexace GATE anotací
  - Pomocí GATE Mímir:
    (Multiparadigm Indexing and Retrieval)
  - https://gate.ac.uk/mimir/

GATE Mímir

- Najdi anotace *Inventor* ve všech dokumentech z roku 2007, kde se slovo „tranzistor“ vyskytuje v *abstraktu*.

```
{Inventor} IN (
{PatentDocument date > 20070000 date < 20080000}
OVER ({Abstract} OVER transistor))
```

- Když připojíme ontologie:
  - Najdi dokumenty kde jsou zmíněny *osoby* narozené v Sheffieldu.

```
{Person sparql = "SELECT ?inst WHERE
{ ?inst :birthPlace
<http://dbpedia.org/resource/Sheffield>}"}
```

On-line demo: http://demos.gate.ac.uk/mimir/

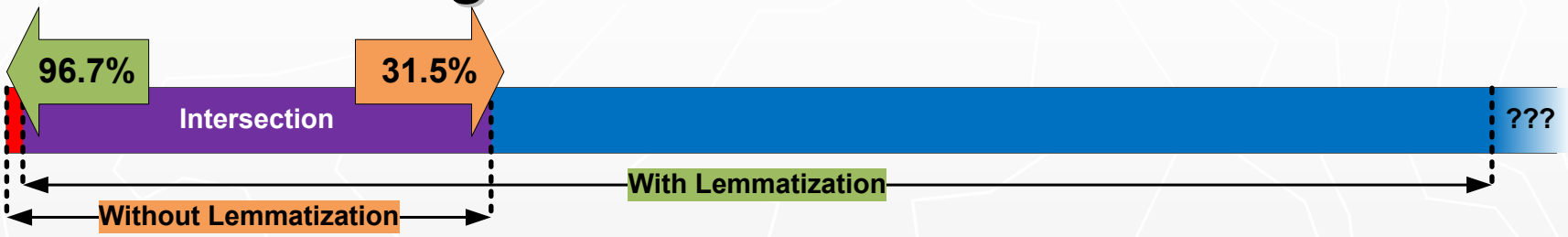
Příklad – Čeština v projektu Khresmoi

- http://khresmoi.eu/
- Databáze lékařských článků a její indexace
  - GATE Mímir
- Anotace lékařských termínů
  - Vzájemné porovnání úspěšnosti různých přístupů díky GATE evaluaci
  - Hierarchický index
  - Jazykově neutrální vyhledávání
- On-line demo (link na požádaní)

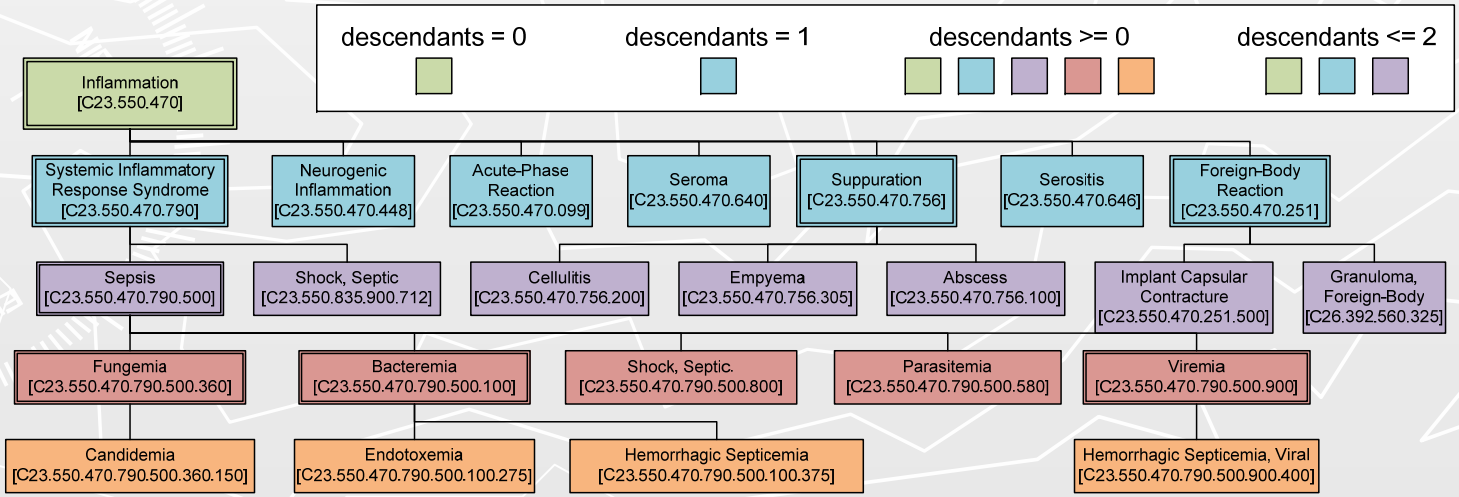
khresmoi
MEDICAL INFORMATION ANALYSIS & RETRIEVAL

Khresmoi

- Evaluace – gazetteer s lemmatizací a bez ní



- Hierarchický index (MeSH)



Děkuji za pozornost!

- ... následuje brainstorming