

Ontology Learning
Ontology Building

Zostavovanie ontológií

Marián Šimko

Ontológia

- Načo?
 - sémantika
 - usudzovanie, odvodzovanie

Ontológia

- Načo?
 - sémantika
 - usudzovanie, odvodzovanie
- Ako?
 - manuálne (doménoví experti)
 - automaticky! (-zovane)

Zostavovanie ontológií

- automatická(-zovaná) tvorba ontológií
- zdroje:
 - slovníky
 - DB
 - text
 - web

Zostavovanie ontológií

- automatická(-zovaná) tvorba ontológií
- zdroje:
 - slovníky
 - DB
 - text
 - web
 - Web

Zostavovanie ontológií

- automatická(-zovaná) tvorba ontológií
- zdroje:
 - slovníky
 - DB
 - text
 - web
 - Web

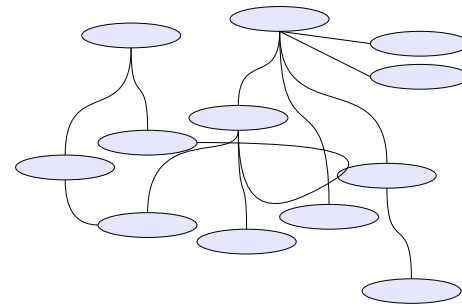
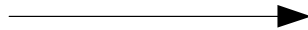
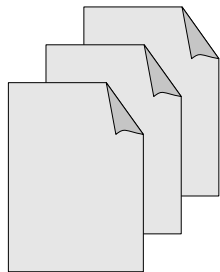
Zostavovanie ontológií

- automatická(-zovaná) tvorba ontológií
- zdroje:
 - slovníky
 - DB
 - text
 - web
 - Web

Zostavovanie ontológií z textu

Zostavovanie ontológií z textu

- metódy a techniky Dolovania v texte

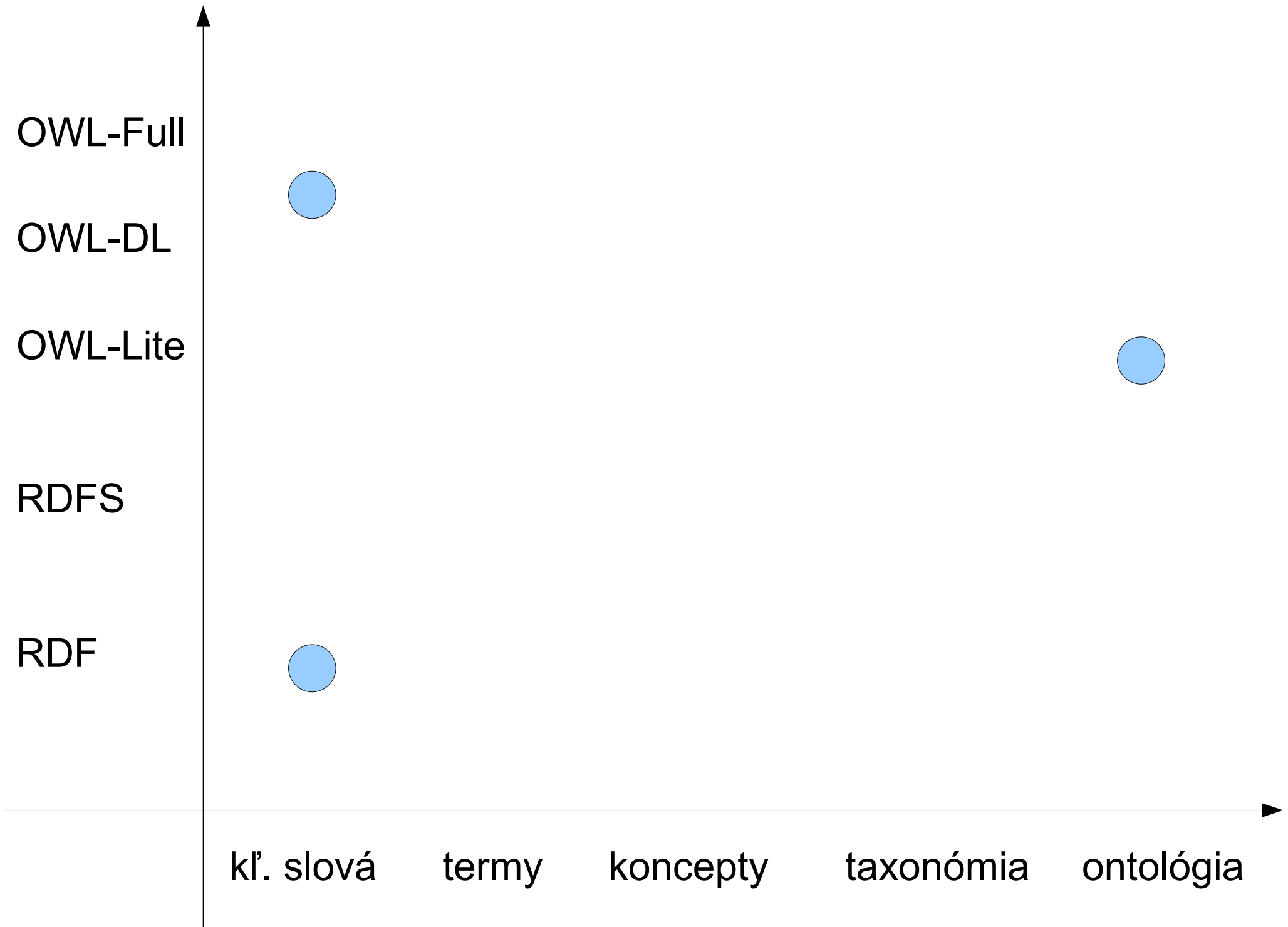


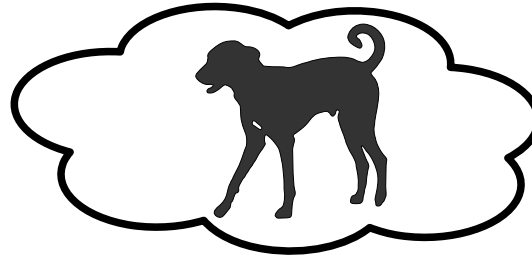
--	--

koncepty

Koncept, pojem

pojem -jmu *m.* forma myslenia odrážajúca podst. vlastnosti predmetov, javov objektívnej skutočnosti: *všeobecný, jedinečný p.*; *myslíme v p-och a súdoch*;
filoz. výsledok rozumového st. poznania skutočnosti (predmetov, javov ap.)





koncept
(pojem, designát)

designácia

sinifikácia

objekt
(vec, denotát)



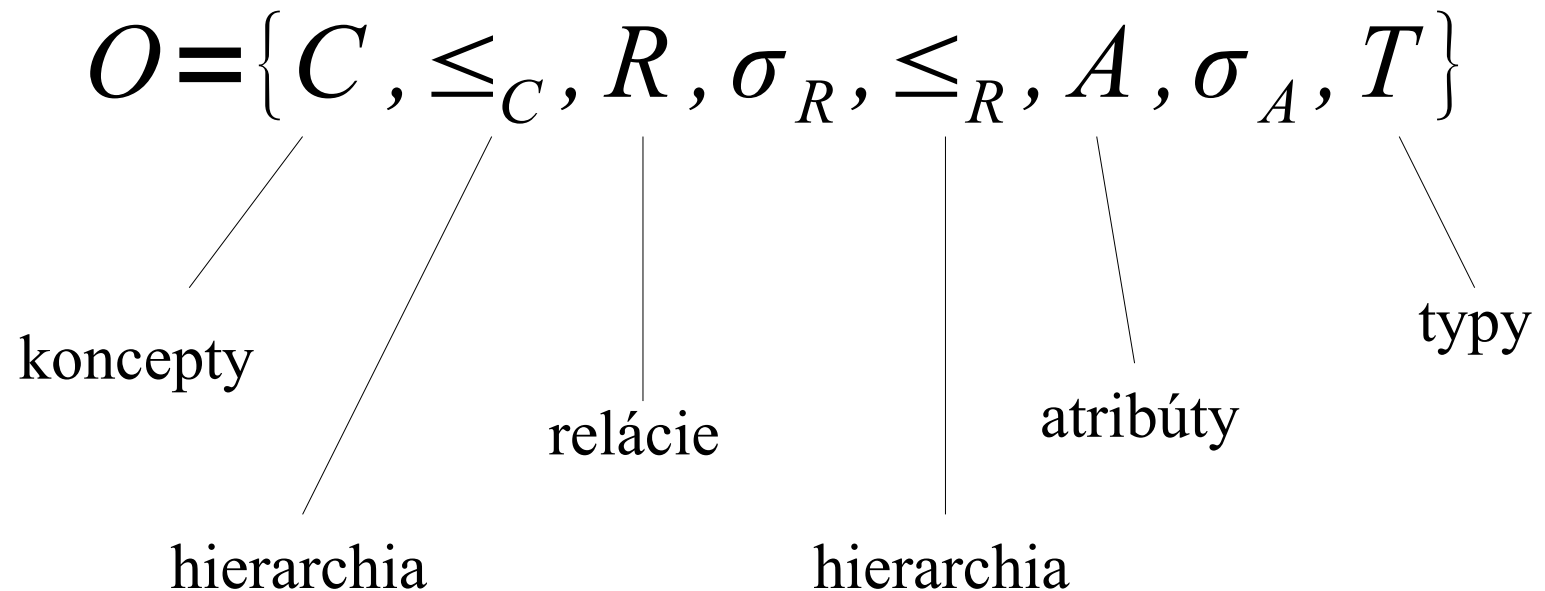
denotácia

pes

symbol
(slovo, designátor)

Ontológia

- Aristoteles: taxonómia
- Teraz: osmica

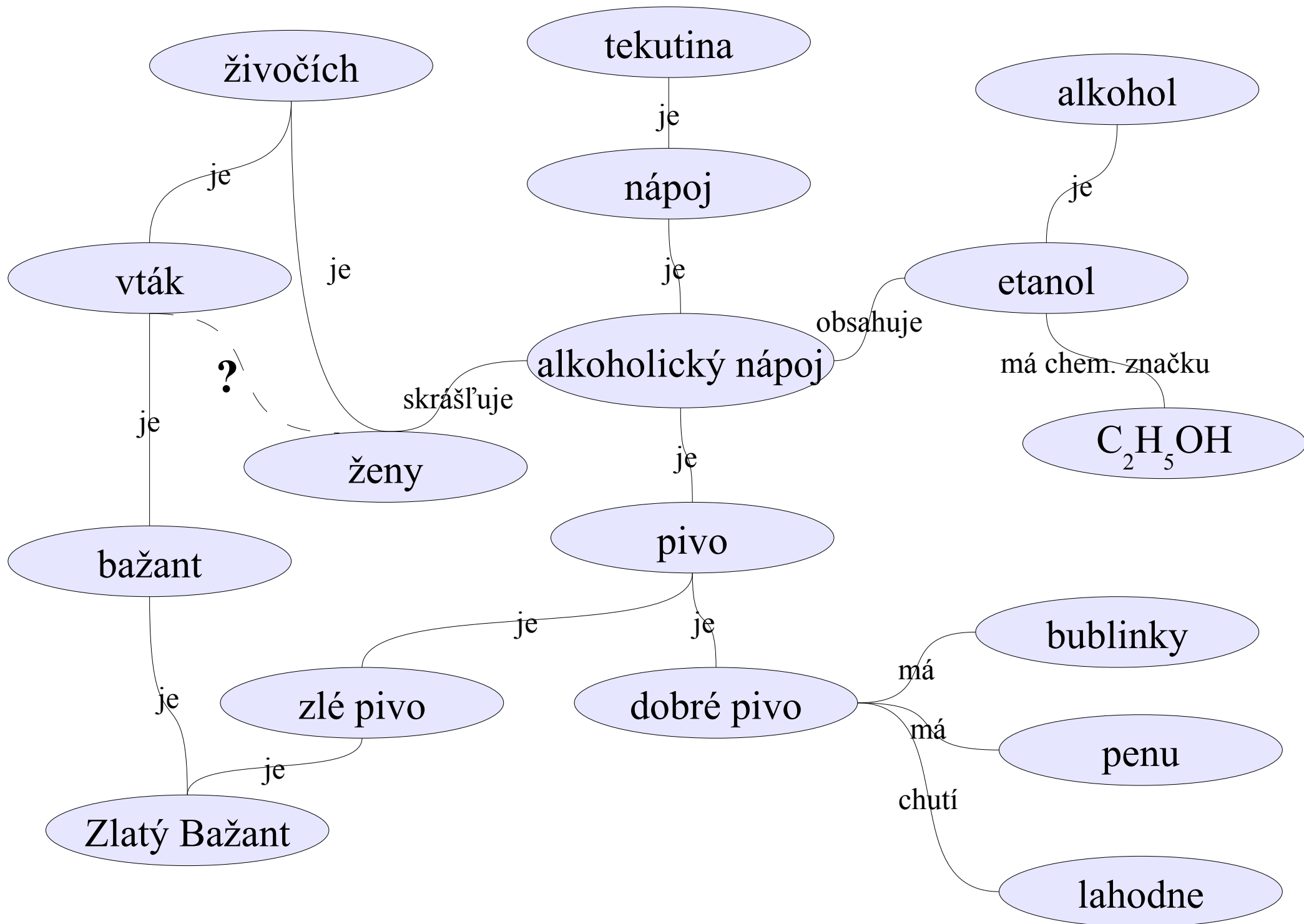


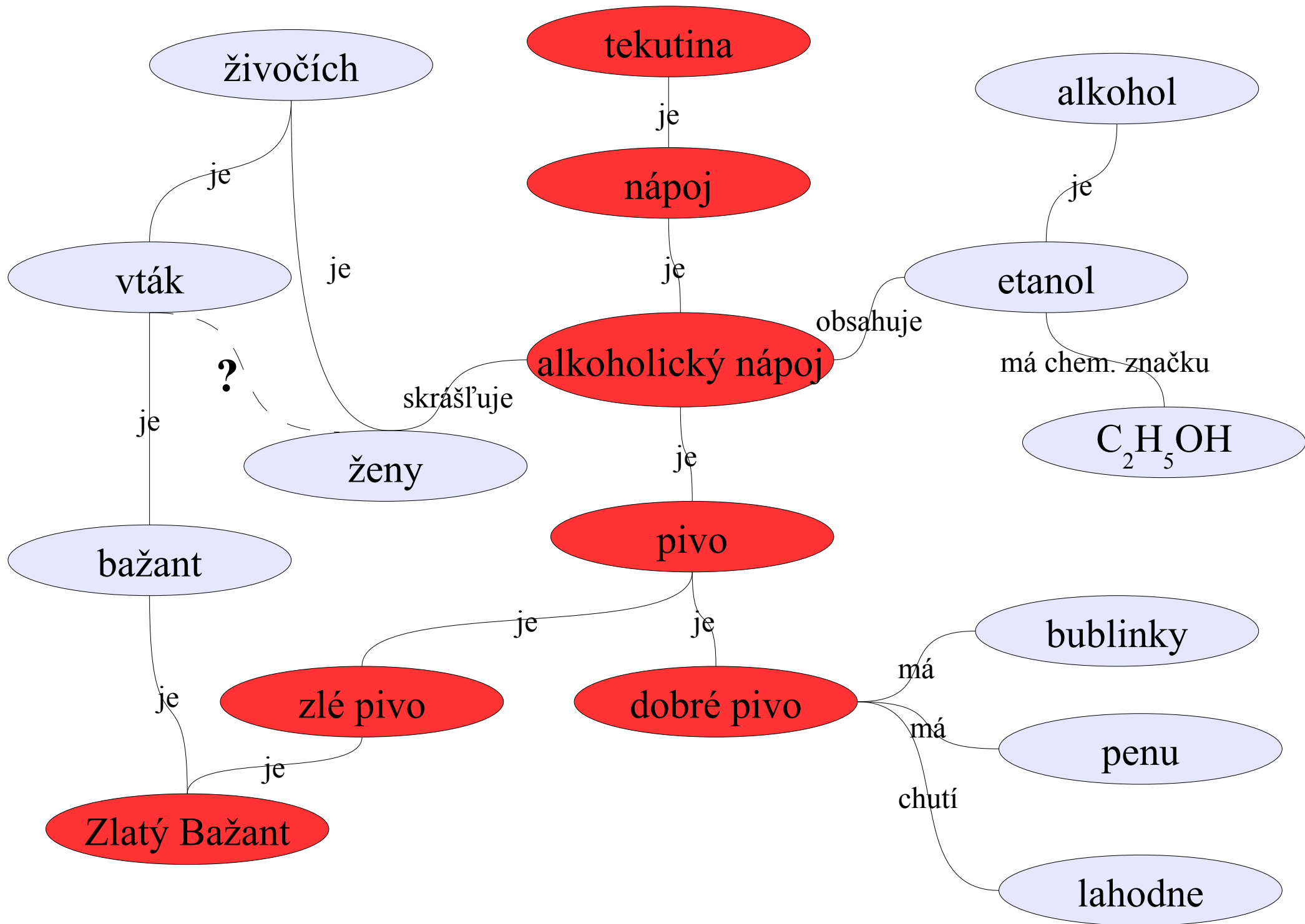
Zostavovanie ontológie z textu

1. Predspracovanie textu
2. Zostrojenie konceptov
3. Tvorba relácií
 - a) Hierarchických
 - b) Nehierarchických
4. Pokročilé úlohy
5. Populácia ontológie

Príklad







1. Predspracovanie textu

- Spracovanie prirodzeného jazyka (NLP)
- Lexikálna analýza
 - Tokenizácia a normalizácia
 - Označenie slovných druhov (POS tagging)
 - Lematizácia, stemovanie

1. Predspracovanie textu

- Kúskovanie (plytké parsovanie)
 - „meaning-bearing unit“
- Syntaktická analýza
 - Syntaktické závislosti
 - „propositional phrases“
 - Pseudo-syntaktické závislosti
 - „a nice city“ -> nice(city)
 - Kolokácie

2. Zostrojenie konceptov

- Extrakcia termov
 - Automatické extrakcia termov
 - ATR – automatic term recognition
 - Identifikácia vlastných mien
 - NER – named entity recognition
 - Spracovanie synonym
- WordNet

3.a. Tvorba hierarchických relácií

- Spracovanie slovníkov
 - MRD – machine readable dictionary
 - Definície na „podnose“
 - Vysoká presnosť :)
 - Doménovo nešpecifické :(
- Lexikálno-syntaktické vzory
 - „*NP such as {NP,}* {(and | or)} NP*“
 - nepokrývajú dostatok vzťahov v texte :(
 - vzťahy skôr medzi slovami ako konceptmi :(

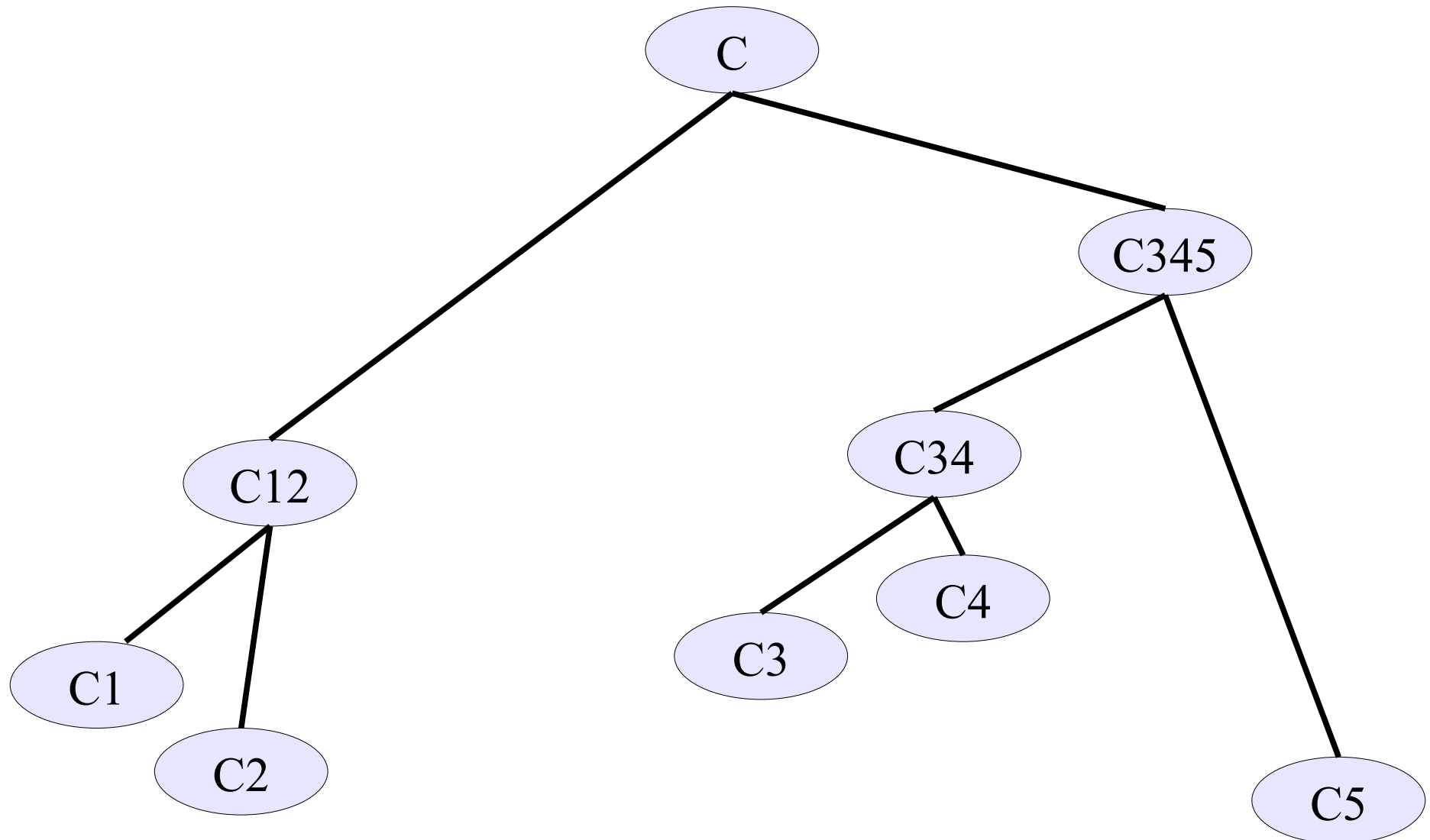
3.a. Tvorba hierarchických relácií

- Analýza textu - tzv. *distributional hypothesis*

„Slová sú si podobné do takej miery, do akej zdieľajú spoločný kontext.“ [Harris, 1998]

- Štatistický prístup
 - Zhlukovanie (divisive clustering, hierarchical clustering)
- Množinový prístup
 - FCA – formal concept analysis

Hierarchické zhľukovanie



3.b. Tvorba nehierarchických relácií

- Najmenej prebádané
- Analýza spoločných výskytov
 - Co-occurrence analysis
- Syntaktické závislosti
- Lexikálno-syntaktické vzory

4. Pokročilé úlohy

- Inštanciácia schém axióm
 - Ktoré množiny konceptov sú disjunktné
 - Ktoré relácie sú tranzitívne
 - Kardinalita relácií
- Generalizácia axióm
 - V závislosti od logického formalizmu

Zhodnotenie

- Ľudský jazyk je stratová kompresia
- Vysoká miera „šumu“
- Konflikt konceptualizácie
 - Každý máme v hlave inú ontológiu
- Stále nízka úspešnosť metód zostavovania ontológie

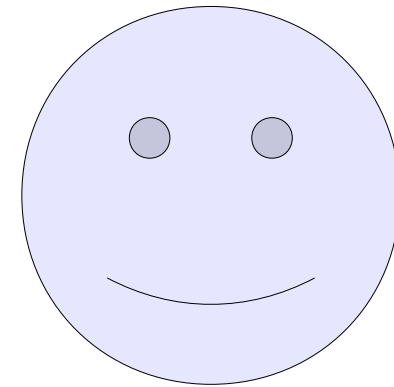
Zostavovanie ontológií z Webu

Zostavovanie ontológií z Webu

- Web ako úložisko poznania ľudstva
 - Zostavovanie ontológií z obsahu (textu)
 - Text mining
 - Analýza topológie Webu
 - Structure mining
 - Sociálne interakcie
 - Tagy, komentáre
 - Záznamy dopytov
- } Usage mining

Zostavovanie ontológií z Webu

- Viac možno nabudúce...



Zdroje a nástroje

- Cimiano, P. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer-Verlag, ISBN: 978-0-387-30632-2. 347p (2006)
- Lucene
 - IR knižnica, indexovanie, vyhľadávanie
 - <http://lucene.apache.org>
- JUNG
 - Grafové algoritmy: PageRank, HITS, etc.
 - <http://jung.sourceforge.net>