PA153 Počítačové zpracování přirozeného jazyka

04 – Sémantika I
(reprezentace lexikálního významu)

Karel Pala, Zuzana Nevěřilová

Centrum ZPJ, FI MU, Brno

23. října 2014

1. Lexikální význam
2. Slovníkové heslo
3. Nalezení významu v kontextu
   - Algoritmy lexikální desambiguace
4. Popis lexikálních významů pro ZPJ
   - Sémantické primitivy
   - Sémantické třídy
   - Teorie prototypů
5. Shrnutí

Lexikální význam

lexikální význam (*lexical meaning*): izolovaný význam slova [Oxford Dictionaries, 2013]

- bez ohledu na význam věty, ve které se slovo nachází
- bez ohledu na gramatické kategorie

jiné významy: gramatický význam, význam slov a význam vět

- *kuře – kuřata*
- *frekvence – kmitočet*
- Pan profesor *běží* na tramvaj. Gepard *běží* za kořistí.

lexikální význam (*lexical meaning*): izolovaný význam slova [Oxford Dictionaries, 2013]

- bez ohledu na význam věty, ve které se slovo nachází
- bez ohledu na gramatické kategorie

jiné významy: gramatický význam, význam slov a význam vět

- *kuře – kuřata*
- *frekvence – kmitočet*
- Pan profesor *běží* na tramvaj. Gepard *běží* za kořistí.

slova kuře a kuřata mají tentýž lexikální význam, ale rozdílný gramatický (singulár, plurál)

frekvence a kmitočet jsou různá slova, která mají tentýž lexikální (i gramatický a dokonce i další) význam

běžet má stejný význam, přestože si představíme celkem jinou činnost (styl, rychlost, terén)

Lexikální forma a lexikální význam

Lexikální jednotka (lexical unit, LU) [Ziková, 2003]:

- reprezentována lexikální formou
- asociována s určitým lexikálním významem
- má určité gramatické vlastnosti (např. tranzitivní sloveso)
- může mít určité pragmatické vlastnosti (např. *já* je pokaždé někdo jiný)

- LU se stejným významem, ale jinou formou
  synonymie (např. šalina, tramvaj, šmirgl)
- LU se stejnou formou, ale jiným významem
  homonymie (např. kolej)

Kde najít informace o lexikálním významu?

Slovník/lexikon/lexikální databáze = soubor lexikálních jednotek (LU)

Slovníky:

- jednojazyčné výkladové
- překladové
- současného jazyka (synonym, zkratek, rýmů . . .)
- terminologické
- historické
- etymologické
- speciální (frekvenční, retrográdní, valenční)
- . . .

strojově čitelné slovníky = machine readable dictionaries

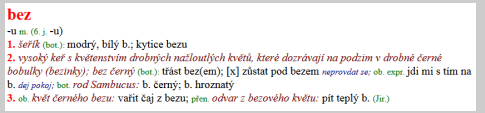
Struktura slovníkového hesla

bez

-u m. (6. j. -u)
1. *šeřík* (bot.): modrý, bílý b.; kytice bezu
2. *vysoký keř s květenstvím drobných nažloutlých květů, které dozrávají na podzim v drobné černé bobulky (bezinky); bez černý* (bot.): třást bez(em); [x] zůstat pod bezem *neprovdat se;* ob. expr. jdi mi s tím na b. *dej pokoj;* bot. *rod Sambucus:* b. černý; b. hroznatý
3. ob. *květ černého bezu:* vařit čaj z bezu; přen. *odvar z bezového květu:* pít teplý b. (Jir.)

- lexikální forma
- gramatické vlastnosti
- definice
- kolokace
- příklady užití
- odvozené lexikální formy (hnízdování)

Struktura slovníkového hesla


bez
-u m. (6. j. -u)
1. *šeřík* (bot.): modrý, bílý b.; kytice bezu
2. *vysoký keř s květenstvím drobných nažloutlých květů, které dozrávají na podzim v drobné černé bobulky (bezinky); bez černý* (bot.): třást bez(em); [x] zůstat pod bezem *neprovdat se;* ob. expr. jdi mi s tím na b. *dej pokoj;* bot. *rod Sambucus:* b. černý; b. hroznatý
3. ob. *květ černého bezu:* vařit čaj z bezu; přen. *odvar z bezového květu:* pít teplý b. (Jir.)

- lexikální forma
- gramatické vlastnosti
- definice
- kolokace
- příklady užití
- odvozené lexikální formy (hnízdování)

Ukázka je ze SSJČ, kde není žádná odvozená forma, ale v SSČ je bezový a bezinka. Pro jiná slova je v SSČ mnohem více odvozených forem: květ, květen, květena, květák, květenství, květina, květináč, květinářka, květinářství

Kolokace jako slovníkové heslo

pevné kolokace: zakopaný pes, devítiocasá kočka, slaměný vdovec, New York, křížem krážem, ad hoc

porušují princip kompozicionality
samostatná slovníková hesla?

v NLP se používá termín multiword expresion (MWE)
je důležité MWE identifikovat, např. pro strojový překlad:

- pevné MWE: zakopaný pes
- vzory: vzít <*někoho*> na hůl

Slovníkové definice a hyperonymie

bez

-u m. (6. j. -u)
1. *šeřík* (bot.): modrý, bílý b.; kytice bezu
2. *vysoký keř s květenstvím drobných nažloutlých květů, které dozrávají na podzim v drobné černé bobulky (bezinky); bez černý* (bot.): třást bez(em); [x] zůstat pod bezem *neprovdat se;* ob. expr. jdi mi s tím na b. *dej pokoj;* bot. *rod Sambucus:* b. černý; b. hroznatý
3. ob. *květ černého bezu:* vařit čaj z bezu; přen. *odvar z bezového květu:* pít teplý b. (Jir.)

Definice pomocí synonym:
bez = šeřík

Definice klasická:
bez = vysoký keř s květenstvím drobných nažloutlých květů. . . [Havránek et al., 1960]

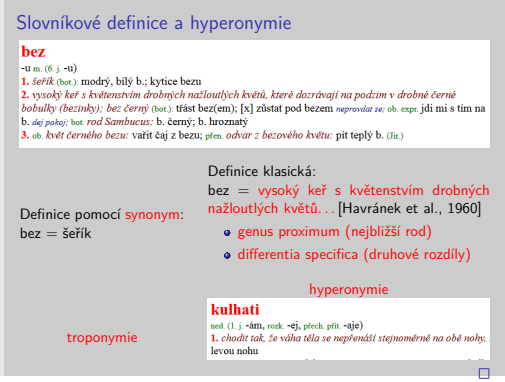
- genus proximum (nejbližší rod)
- differentia specifica (druhové rozdíly)

hyperonymie

troponymie

kulhati

ned. (1. j. -ám, rozk. -ej, přech. přít. -aje)
1. *chodit tak, že váha těla se nepřenáší stejnoměrně na obě nohy,*
levou nohu



klasické slovníkové heslo vyžaduje pro porozumění aspoň nějakou znalost jazyka, tudíž je pro počítačové zpracování neúplné

Nalezení významu v kontextu

někdy (ve skutečnosti velmi často) jen se znalostí lexikálního významu nevystačíme
$\Rightarrow$ je třeba znát kontext

lexikální desambiguace (Word Sense Disambiguation)
funkce: $(w, c) \rightarrow s$

- $w \in \mathcal{W}$ – množina slov
- $c \in \mathcal{C}$ – množina kontextů
- $s \in \mathcal{S}$ – množina významů

Naivní Leskův algoritmus: list (SSJČ) [Lesk, 1986]

1. jeden ze základních orgánů rostlin, zprav. do plochy rozšířený a velmi různých tvarů; lupen: kaštanový, dubový, javorový l.; velký, malý l.; drobné listy borůvčí; široké listy lip; zelné listy; fíkový l., přen. (ve výtv. dílech) jeho zpodobení zakrývající ohanbí, jednání ap. věcně něco zastírající;
2. kniž. a nář. listí: svěžím listem zalesklo se habří (Jir.); stromy obalily se listem (Něm.)
3. kus papíru čtyřúhelníkového tvaru, zprav. určený k psaní, tisku ap.: sešit o 24 listech; titulní l. v knize; l. pergamenu; cyklus grafických listů; její duše je nepopsaný l. (kniž.) nemá zkušenosti; zpívat, hrát přímo z listu z notového partu bez cvičení; . . .
4. kniž. a zast. dopis, psaní: zalepený, zapečetěný l.; otevřený l.; veřejný, osobní l.; listy Jana Nerudy; hist. opovědný, odporný, výhostní l.; církev. apoštolský, pastýřský l. provolání, výzva papeže, biskupa
5. úřední listina o něčem svědčící, k něčemu opravňující: rodný, domovský (dř.), oddací, úmrtní l.; výuční, živnostenský l.; odběrní, dodací l.; nákladní l.; záruční, zástavní l.; vůdčí l. (dř.) řidičský

Naivní Leskův algoritmus: vstup

Ještě lepším řešením by bylo vydat se evropskou cestou: zbavit se úvěrů bez zodpovědnosti dlužníka a rozvinout systém financí založený na zástavních *listech*, jako jsou německé Pfandbriefe.

{a, bez, by, být, cesta, dlužník, dobrý, evropský, finance, jako, ještě, německý, rozvinout, řešení, se, systém, úvěr, vydat, založený, zástavní, zbavit, zodpovědnost}

Naivní Leskův algoritmus

{a, bez, by, být, cesta, dlužník, dobrý, evropský, finance, jako, ještě, německý, rozvinout, řešení, se, systém, úvěr, vydat, založený, zástavní, zbavit, zodpovědnost}

1:{a, borůvčí, dílo, do, drobný, dubový, fíkový, javorový, jeden, jednání, kaštanový, lípa, lupen, malý, ohanbí, orgán, plocha, rostlina, rozšířený, různý, široký, tvar, věčně, velký, velmi, . . . }
2:{habří, listí, obalit, se, strom, svěží, zalesknout}
3:{bez, být, cvičení, cyklus, čtyřúhelníkový, dnes, duše, grafický, hráč, hrát, jeden, jeho, jiný, k, karta, kniha, který, kus, mít, mluvit, něco, notový, o, obrátit, on, padat, papír, part, pergamen, popsaný, přímo, psaní, ruka, se, sešit, situace, souhrn, štěstí, tvar, určený, v, tisk, titulní, z, . . . }
4:{apoštolský, biskup, dopis, Jan, Neruda, odporný, opovědný, osobní, otevřený, papež, pastýřský, provolání, psaní, veřejný, výhostní, výzva, zalepený, zapečetěný}
5:{dodací, domovský, k, kniha, listina, nákladní, něco, o, odběrní, oddací, opravňující, průkaz, pozemkový, rodný, řidičský, svědčící, úmrtní, úřední, vůdčí, výuční, záruční, zástavní, živnostenský}

Naivní Leskův algoritmus

Ještě lepším řešením by bylo vydat se evropskou cestou: zbavit se úvěrů bez zodpovědnosti dlužníka a rozvinout systém financí založený na zástavních *listech*, jako jsou německé Pfandbriefe.
{a, bez, by, být, cesta, dlužník, dobrý, evropský, finance, jako, ještě, německý, rozvinout, řešení, se, systém, úvěr, vydat, založený, zástavní, zbavit, zodpovědnost}

$D_1 = \{\text{a}\}$
$D_2 = \{\text{se}\}$
$D_3 = \{\text{bez, být, se}\}$
$D_4 = \{\}$
$D_5 = \{\text{zástavní}\}$

kus papíru čtyřúhelníkového tvaru, zprav. určený k psaní, tisku ap. . . .

Ještě lepším řešením by bylo vydat se evropskou cestou: zbavit se úvěrů bez zodpovědnosti dlužníka a rozvinout systém financí založený na zástavních *listech*, jako jsou německé Pfandbriefe.
{a, bez, by, být, cesta, dlužník, dobrý, evropský, finance, jako, ještě, německý, rozvinout, řešení, se, systém, úvěr, vydat, založený, zástavní, zbavit, zodpovědnost}

$D_1 = \{a\}$
$D_2 = \{se\}$
$D_3 = \{bez, být, se\}$
$D_4 = \{\}$
$D_5 = \{zástavní\}$

kus papíru čtyřúhelníkového tvaru, zprav. určený k psaní, tisku ap. ...

Naivní L. algoritmus určil, že význam slova *list* v uvedené větě je 3. Je to spíš náhoda podpořená tím, že u významů 1 a 3 v SSJČ také nejvíc textu.

Vylepšené verze L. algoritmu některá slova nepočítají, přidávají slovům váhy (např. pomocí TF-IDF), zohledňují vzdálenost od desambiguovaného slova

Slabiny Leskova algoritmu

slovníkové definice a příklady užití

WSD založené na metodách strojového učení [Yarowsky, 1995]

1. stanovit význam u pevných kolokací (ručně nebo ze slovníku)
   obrátit list (list:3), živnostenský list (list:5), ...
2. iterativně zjistit další kolokace
   kopie (živnostenského listu) $\rightarrow$ kopie oddacího listu (list:5)
3. opakovat, dokud desambiguované množiny nepřestanou narůstat

Algoritmus natrénovaný na obecném korpusu je použitelný na dalších textech.

Slabiny WSD

$(w, c) \rightarrow s$

- $w \in \mathcal{W}$ – množina slov
- $c \in \mathcal{C}$ – množina kontextů
- $s \in \mathcal{S}$ – množina významů

Všechny algoritmy WSD závisejí na inventáři a popisu významů.

Kolik významů má slovo *list*?

- SSJČ: 8
- SSČ: 6
- Slovník českých synonym: 4
- Český WordNet: 9

$(w, c) \rightarrow s$

- $w \in \mathcal{W}$ – množina slov
- $c \in \mathcal{C}$ – množina kontextů
- $s \in \mathcal{S}$ – množina významů

Všechny algoritmy WSD závisejí na inventáři a popisu významů.

Kolik významů má slovo *list*?

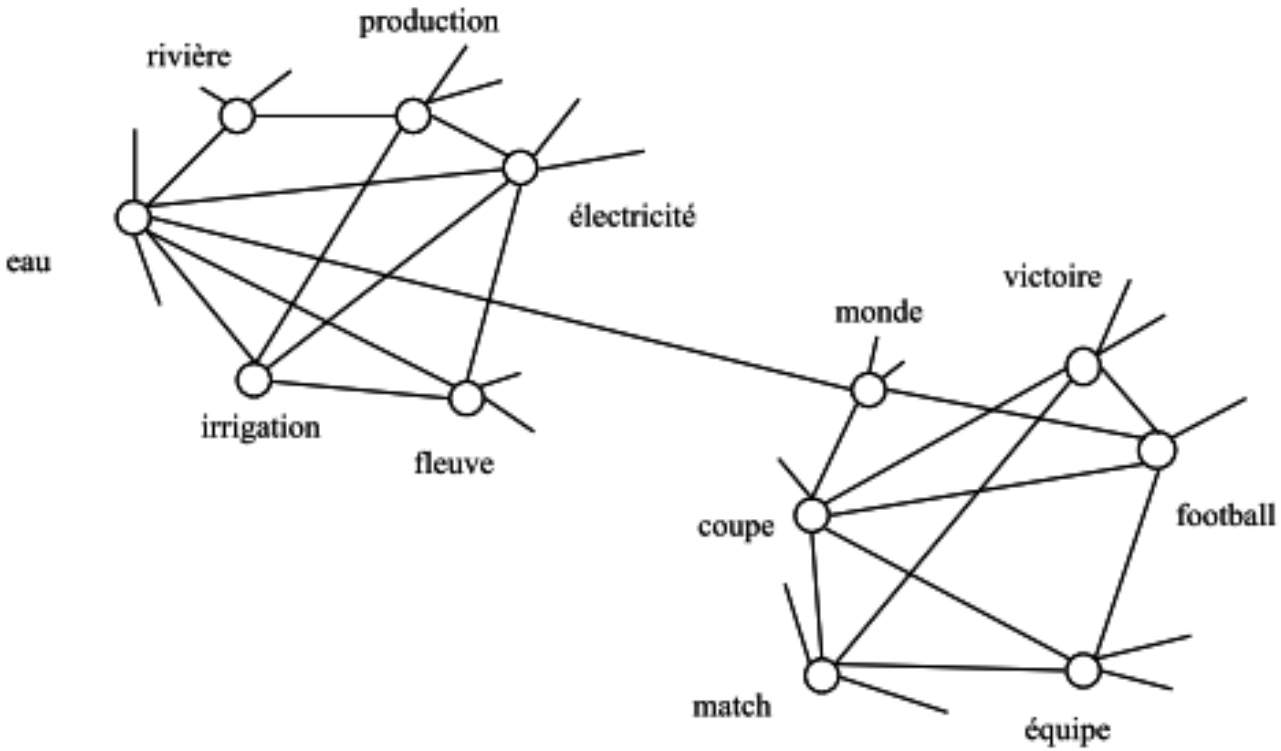
- SSJČ: 8
- SSČ: 6
- Slovník českých synonym: 4
- Český WordNet: 9

Leskův a. je jednoduchý i ve svých pokročilejších verzích, zajímavý algoritmus nabídl [Yarowsky, 1995]. Jde o a. strojového učení, kdy se v prvním průchodu určí kolokace, které naprosto jistě souvisejí s konkrétním významem slova. V dalších průchodech se vypočítávají další slova, která signalizují konkrétní význam slova.

WSD nebo WSD

Algoritmy, které nepočítají s pevným inventářem významů, jen s kontextem:

Word Sense *Discrimination*



[Véronis, 2004]

Komponentová analýza (*Componential analysis*)

= popis významů slov pomocí množiny sémantických rysů (primitiv), které jsou buď přítomny, nebo nepřítomny, nebo irelevantní pro daný význam:

- muž = +HUMAN +ADULT +MALE
- žena = +HUMAN +ADULT -MALE
- chlapec = +HUMAN -ADULT +MALE
- batole = +HUMAN -ADULT ±MALE

[Katz and Fodor, 1963] a [Bierwisch, 1971]

Komponentová analýza (Componential analysis) I

označení	popis	příklad
T	tempus, čas	den, rok, leden, soumrak
L	locus, místo	dům, chrám, světadíl, břeh
BYT	bytost	víla
HUM	člověk	strejda, rada, bača
ANIM	zvíře	pes, slon, velbloud
PLANT	rostlina	strom, kosatec
QUA	vlastnost	nespokojenec, povýšenec + HUM
FEN	fenomén	úkaz, zázrak
ENT	entita	protiklad, argument
OBJ	objekt, předmět	stůl, krb, ale i dům (OBJ + L)

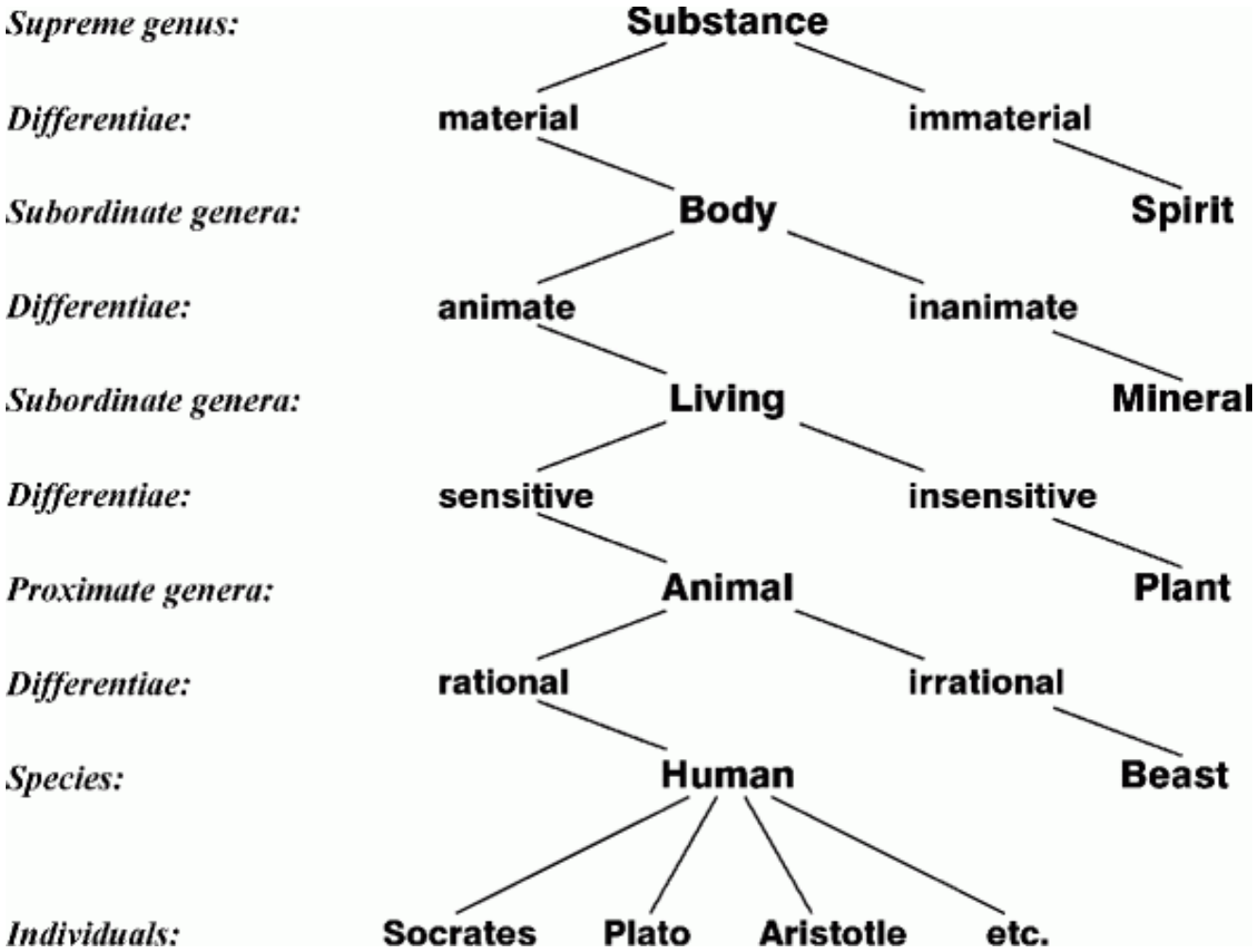
Komponentová analýza (Componential analysis) II

označení	popis	příklad
INF	informace	telefonát, článek, vzkaz
EMO	emoce	cit, radost, strach, neklid, úsměv
INS	instrument, nástroj	nůž, šíp hřeben
MACH	stroj, aparát, zařízení	počítač
PROC	proces	zážeh, postup, pokrok
MOT	pohyb	běh, let, pád
AKT	aktivita, činnost	boj, odboj, příchod
MAT	materiál	hlína, dřevo
BP	část těla (body part)	prst, krk
ORG	organizace, instituce	vláda

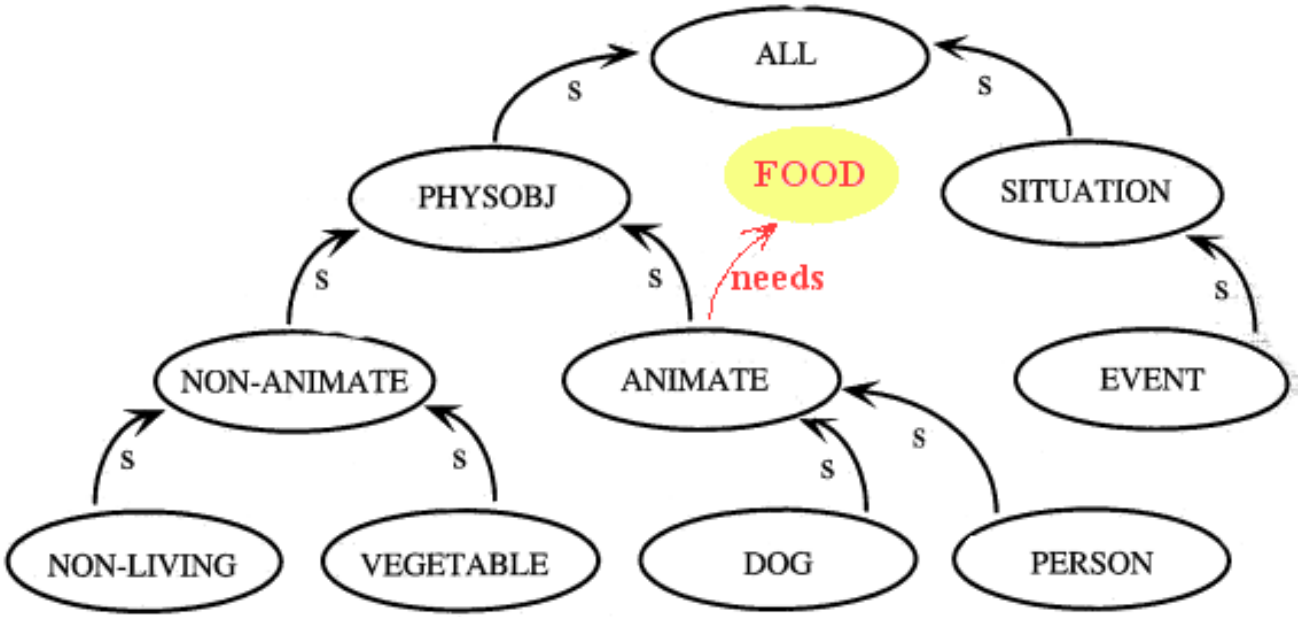
Sémantické třídy

= skupiny slov, která sdílejí určitý sémantický rys
obratlovec – savec – šelma – psovitá šelma – pes – pudl – trpasličí pudl
taxonomie, hierarchie tříd

Sémantické třídy, Porfyriův strom



Sémantické třídy, sémantické sítě, odvozování



Sémantické sítě

WordNet (Princeton WordNet, PWN) – lexikální síť

- původně nástroj k ověření teorie o uspořádání lidské paměti (G. A. Miller, od r. 1985)
- počítačově dobře zpracovatelný zdroj informací o významech slov a vztazích mezi významy [Fellbaum, 1998]
- jednotkou je synonymická řada (synonymical set, synset)
- synsety jsou spojeny relacemi:
  - hyperonymie/hyponymie: vůz, automobil – dodávka
  - holonymie/meronymie (part of, member of): vůz, automobil – tlumič; orchestr – houslista
  - troponymie: šeptat – mluvit
  - near-antonym: den – noc
  - odvození: velikost – velký
- slovní druhy: substantiva, adjektiva, verba, adverbia

WordNet

angličtina: PWN (117 tis. synsetů)
projekty EuroWordNet (angličtina + holandština, italština, španělština, němčina, francouzština, čeština, estonština)

- ILI - InterLingual Index
- Top Ontology (63 kategorií)
- Base Concepts

projekty (BalkaNet: bulharština, čeština, rumunština, řečtina, srbština, turečtina), při kterých vznikají wordnety pro další jazyky, koordinátorem databází je Global WordNet Association (GWA)
současný český W.: 28 tis. synsetů

WordNet není jediný

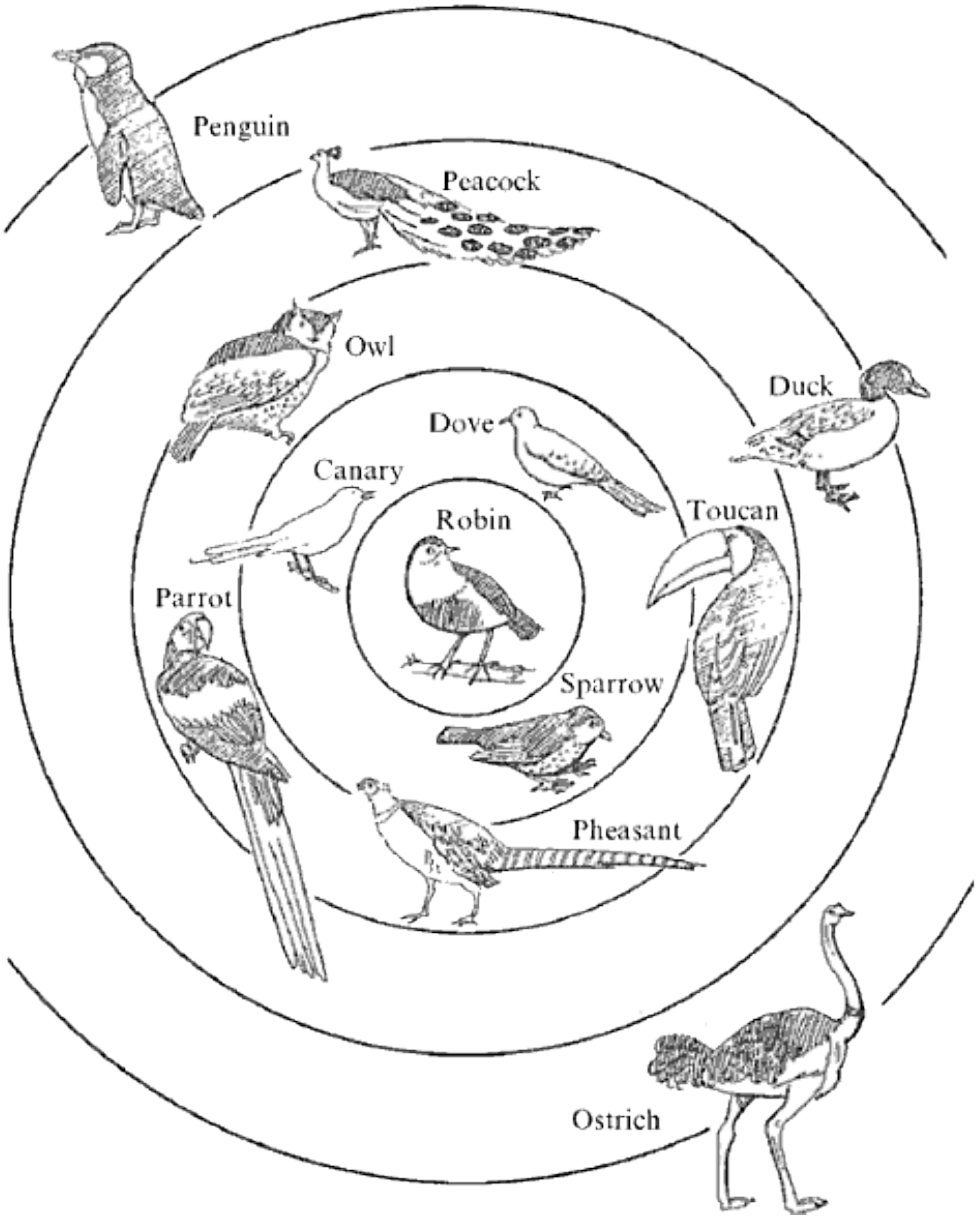
Ontologie = explicitní specifikace sdílené konceptualizace

- firemní o.
- všeobecné o. SUMO/MILO (Suggested Upper Merged Ontology, MId-Level Ontology)
- common sense o. ConceptNet

Ontologie a datové formáty (ontologické jazyky)

- predikátová logika 1. řádu a rozšíření
- Rodina KIF (Knowledge Interchange Format)
- Rodina RDF (Resource Description Framework), „jazyky sémantického webu“: RDF, RDFS, OWL, DAML

Teorie prototypů: co je to ptáček?



Aitchison, 2003 in [Goddard, 2011]

Teorie prototypů

E. Rosch dokázala, že lidé uvažují o vlastnostech třídy jako o vlastnostech typického zástupce třídy.
t. prototypů se uplatňuje v popisu typických situací (rámce, skripty)
vzdálenost mezi koncepty: židle *je víc nábytek než* sporák

Shrnutí

gramatika	slovní druh, gramatické kategorie
syntax	větný člen
sémantika	sémantická třída

popis lexikálního významu:

- pro uživatele jazyka: slovníky
- pro počítačové programy: specializované zdroje (sém. rysy, ontologie, prototypy)

rozlišení lexikálního významu:

- pro uživatele jazyka: číslo významu
- pro počítačové programy: WSD, vzdálenost mezi koncepty

Odkazy I


Bierwisch, M. (1971).
On classifying semantic features.
In M. Bierwisch, K. E. H., editor, *Progress in Linguistics*, pages 27–50. Mouton.


Fellbaum, C. (1998).
WordNet: An Electronic Lexical Database (Language, Speech, and Communication).
The MIT Press.
Published: Hardcover.


Goddard, C. (2011).
Semantic Analysis: A Practical Introduction.
Oxford Textbooks in Linguistics. Oxford University Press.

Odkazy II

Havránek, B. et al. (1960).
Slovník spisovného jazyka českého (Dictionary of Written Czech, SSJČ).
Academia, Praha, 1st edition.
electronic version, created in the Institute of Czech Language, Czech Academy of Sciences Prague in cooperation with Faculty of Informatics, Masaryk University Brno.

Katz, J. and Fodor, J. (1963).
The structure of a semantic theory.
Language, (39):170–210.

Odkazy III


Lesk, M. (1986).
Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone.
In *Proceedings of the 5th annual international conference on Systems documentation*, SIGDOC '86, pages 24–26, New York, NY, USA. ACM.


Oxford Dictionaries (2013).
lexical meaning. Oxford Dictionaries.
online.
`http://oxforddictionaries.com/definition/english/lexical-meaning` (accessed October 03, 2013).


Véronis, J. (2004).
Hyperlex: Lexical cartography for information retrieval.
In *Computer Speech and Language: Special Issue on Word Sense Disambiguation*, page 23.

Odkazy IV

Yarowsky, D. (1995).
Unsupervised word sense disambiguation rivaling supervised methods.
In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ACL '95, pages 189–196, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ziková, M. (2003).
Současný český jazyk: Tvoření slov.
online.
`http://www.phil.muni.cz/cest/lide/zikova/CJA009_1.rtf`
(accessed October 03, 2013).