Výpočetní sémantika a základní sémantické struktury

Aleš Horák

E-mail: hales@fi.muni.cz
http://nlp.fi.muni.cz/nlp_intro/

Obsah:

- Významy slov a významové vztahy
- Slovníky a specializované lexikony
- Výpočetní sémantika

Významy slov, polysemie

Slovo:

- slovní tvar (*wordform*) – slovo zapsané v textu
- lemma/základní tvar – slovo v indexové/citační podobě (nominativ, singulár, ...)
  váže se na lexikální význam
- lemma i slovní tvar může mít víc významů (*word sense*):
  (pozor na rozdíl *význam jako meaning* a *význam jako sense*)

  ... musel rozhodčí napomínat za vzteklé mlácení raketou ...
  ... cvičně odpálila balistickou raketu středního doletu, která je ...

  ... vystoupení v latinsko-amerických tancích na Vašich kulturních akcích ...
  Při nácviku brodění totiž v tancích navlhly kabely a vojáci je ...

Významy slov, polysemie

Slovo, které má více významů se označuje jako:

- polysémní – významy slova spolu něčím souvisí

  ... mnozí z nich měli v očích slzy ...
  ... zase šlápnutí na kuří oko voličů ...
  ... osmažená vejce na volská oka pokrájená ...
  ... Technologie Jestřábí oko spolehlivě určí, zda byl míček dobrý...

- homonymní – píší se stejně, ale jejich významy spolu nesouvisí
  (může být *homografní* nebo *homofonní* – bít/být)

  ... azuro na obloze, zelená travička pod nohama...
  Jednou z nejslavnějších profesionálních traviček se stala Locusta

  ... zajišťujeme kompletní zákaznický servis ...
  ... Broušený servis , skutečný domácí postrach, který se dědí ...
  ... reklamace zboží v autorizovaném servisu ...
  ... Hingisová sice hned prohrála servis , ale z 0:1 otočila ...

Významy slov, polysemie

Některé typy polysemie jsou systematické:

- budova ↔ organizace ↔ osoby

  ... Nemocnice byla postavena v listopadu 1873 ...
  ... Nemocnice údajně dluží členům asociace 1,5 miliardy ...
  ... Prachatická nemocnice ošetřila také 19 lehce zraněných ...

- viz metonymie – autor ↔ dílo, strom ↔ ovoce

  korigovali text Hovorů proto, že tu bylo více Čapka a méně autentického Masaryka.
  o tom hovořil ve své knize už Karel Čapek ...

  ... u hrázek byla tehdy taková silná hruška ...
  ... tam, kde je na hrušce stopka, ...

Zeugma test na polysemii:

- *Kdo rád stráví silvestrovskou noc při dunění petard?*
- *Pak se však Mach pokusil strávit příliš velké sousto.*
- → *Kdo rád stráví silvestrovskou noc a příliš velké sousto při dunění petard?*

Word Sense Disambiguation

správné určení významu – word sense disambiguation

- WSD má vliv na:
  - vyhledávání informací (určení indexového lemmatu)
  - strojový překlad ("bat" → "netopýr" | "pálka")
  - výslovnost při řečové syntéze
    (angl. "bass [beis]" – bas/basa, "bass [bæs]" – okoun
    čes. "baby [babi]" – mn.č. od baba, "baby [beibi]" – dítě, z angl.)
- klasifikační úloha vztažená k nějakému katalogu významů (sense inventory), např. WordNet
  úspěšnost záleží na vlastnostech katalogu, např. granularita
  nejlepší systémy dosahují cca 60 % pro jemné rozlišení významů a 80 % pro hrubé rozlišení (*fine-grained* × *coarse-grained*)
- klasifikace určuje kontexty odpovídající jednotlivým významům různé metody, od slovníkových po zcela automatické
- bez katalogu je odpovídající úkol word sense induction, určení významů slova podle shluků jeho kontextů

Word Sense Disambiguation – porovnání kontextů

tank/tanec

czes2 freqs = 10,520 | 12,826

tank 6.0 4.0 2.0 0 -2.0 -4.0 -6.0 tanec

coord	503	1,538	1.40	3.10
dělostřelectvo	40	0	9.2	--
peso	14	0	8.6	--
transportér	23	0	8.2	--
houfnice	9	0	8.2	--
pěchota	32	0	8.1	--
kanon	6	0	7.1	--
buldozer	5	0	7.1	--
samopal	7	0	6.4	--
kulomet	6	0	6.2	--
vrtulník	22	0	6.2	--
dělo	26	0	5.8	--
letadlo	60	0	5.7	--
muzika	0	15	--	5.4
rytmus	0	17	--	5.5
kroj	0	5	--	5.5
zábava	0	35	--	5.8
aerobik	0	9	--	6.2
buben	0	12	--	6.4
balet	0	16	--	6.6
šerm	0	9	--	6.9
hudba	0	267	--	7.0
pantomima	0	15	--	7.9
píseň	0	243	--	8.0
zpěv	0	177	--	9.2
poslech	0	147	--	9.8

post_verb	342	498	1.20	1.30
útočit	6	0	4.2	--
vyrábět	5	0	1.9	--
potřebovat	6	0	0.8	--
začít	0	6	--	0.2
patřit	0	18	--	1.3
věnovat	0	7	--	1.3
hrát	0	18	--	1.3
pokračovat	0	8	--	1.3
představit	0	8	--	1.6
začínat	0	12	--	2.0
pomáhat	0	6	--	2.1
vycházet	0	9	--	2.1
bavit	0	6	--	3.2
předvést	0	7	--	3.3
zahrát	0	19	--	4.5

a_modifier	3,218	5,764	1.70	2.20
modernizovaný	65	0	9.0	--
sovětský	318	0	8.4	--
vyprošťovací	29	0	8.0	--
zničený	38	0	7.4	--
zastaralý	24	0	6.9	--
mostní	15	0	6.9	--
Wittmannův	11	0	6.8	--
lehký	103	5	6.9	2.4
moderní	40	238	4.5	7.0
latinskoamerický	8	90	5.8	8.7
povinný	7	91	3.3	6.9
originální	0	69	--	6.8
společenský	0	144	--	6.8
lidový	0	157	--	7.0
dvořákův	0	38	--	7.2
irský	0	69	--	7.3
country	0	77	--	7.8
výrazový	0	50	--	7.8
scénický	0	63	--	8.0
dobový	0	104	--	8.1
rituální	0	62	--	8.2
slovanský	0	104	--	8.2
hříšný	0	92	--	8.7
břišní	0	329	--	10.3
orientální	0	404	--	10.6

Synonyma

Dvě slova jsou synonyma, když jsou vzájemně zaměnitelná v kontextech:

- notebook ↔ laptop
- statečný ↔ odvážný
- chlapec ↔ hoch

většina synonym ale není zaměnitelná ve všech kontextech:

- *Samotný prožitek doteku pak má své kouzlo.*
  Samotný prožitek doteku pak má svůj půvab.
- *Učení nových útočných i obranných pohybů a kouzel.*
  Učení nových útočných i obranných pohybů a půvabů.

Synonymie je tedy vazba mezi významy slov, ne mezi slovy

Antonyma

totéž platí pro antonymii – slova opačného významu nebo stupně vlastnosti:

- tmavý × světlý
- rychle × pomalu
- dovnitř × ven

kontextově jsou antonyma velice podobná synonymům!

tmavý/světlý czes2 freqs = 8,960 | 8,127

tmavý	6.0	4.0	2.0	0	-2.0	-4.0	-6.0	světlý

subj_byt	141	106	8.20	7.40
papír	5	0	1.5	--
obrázek	5	0	1.4	--
obraz	6	0	1.4	--
noc	4	0	1.3	--
barva	4	9	0.9	2.1

modifies	7,316	6,019	5.60	5.60
brýle	205	0	8.8	--
pečivo	51	0	7.3	--
mrak	42	4	6.8	3.5
hnědák	54	7	7.8	5.1
oblek	153	23	8.6	6.0
chléb	63	11	7.2	4.8
pleť	504	123	10.0	8.1
kalhoty	66	19	7.2	5.6
bunda	39	11	6.7	5.1
skvrna	129	38	8.2	6.6
pruh	63	28	6.8	5.7
barva	430	301	7.5	7.1
vlas	325	226	8.5	8.0
chloupek	27	28	6.6	6.9
odstín	150	180	8.3	8.7
pivo	88	138	6.4	7.1
dřevo	41	73	5.8	6.7
ležák	32	68	7.0	8.4
jíška	17	58	6.2	8.2
okamžik	0	118	--	6.7
stezka	0	105	--	7.3
Karolína	0	54	--	7.8
výjimka	0	247	--	8.0
výška	0	343	--	8.1
zítřek	0	239	--	9.7

Hyperonyma a hyponyma

Význam slova w_i je hyperonymum (hyponymum) významu slova u_j, pokud w_i je obecnější (specifičtější):

- kobra je hyponymum slova had
- stroj je hyperonymum bagr

jiné označení:

- slova nadřazená/podřazená (*superordinate*/*subordinate*)
- z logického pohledu u_j je hyponymum $w_i \Leftrightarrow$
  - extenzionálně – $class(u_j) \subset class(w_i)$
  - vyplývání – $property(x, u_j) \Rightarrow property(x, w_i)$
- hypero/hyponymie je obvykle tranzitivní

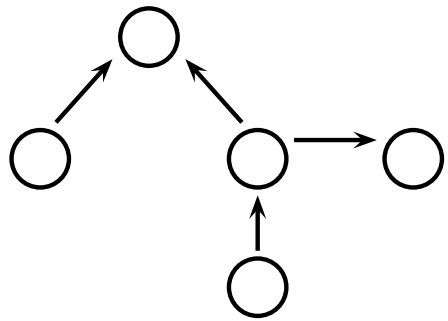
  u_j *hyponymum* $w_i \wedge w_i$ *hyponymum* $v_k \Rightarrow u_j$ *hyponymum* v_k

u sloves podobná relace troponymie – chodit/pochodovat

Sémantické sítě

sémantické sítě – reprezentace faktových znalostí (pojmy + vztahy)

- vznikly kolem roku 1960 pro reprezentaci významu anglických slov
- znalosti jsou uloženy ve formě grafu



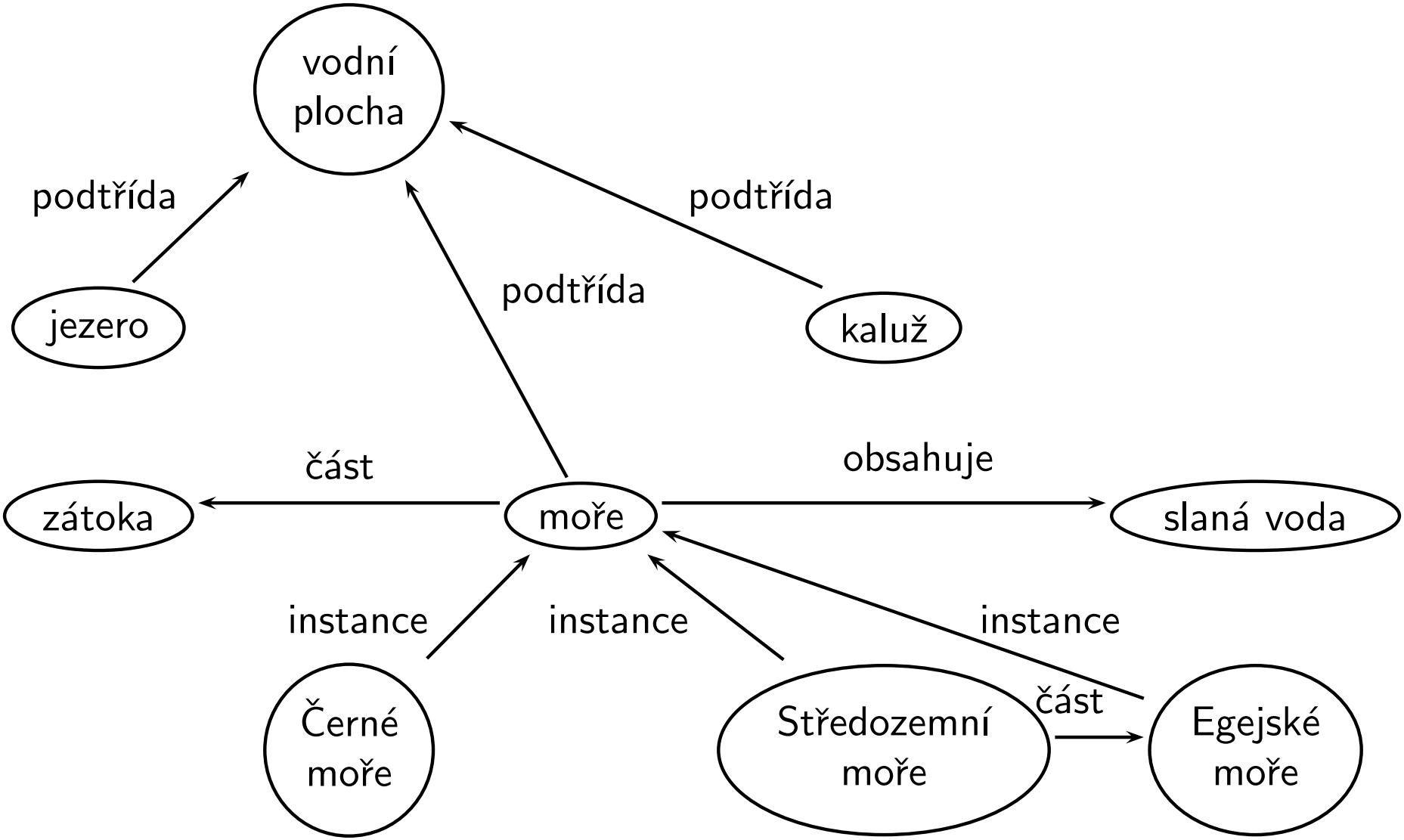
pojmy (objekty, třídy)

vztahy

- nejdůležitější vztahy:
  - podtřída (*subclass*, *is-a*) – vztah mezi třídami
  - instance – vztah mezi konkrétním objektem a jeho rodičovskou třídou

jiné vztahy – část (has-part), barva, ...

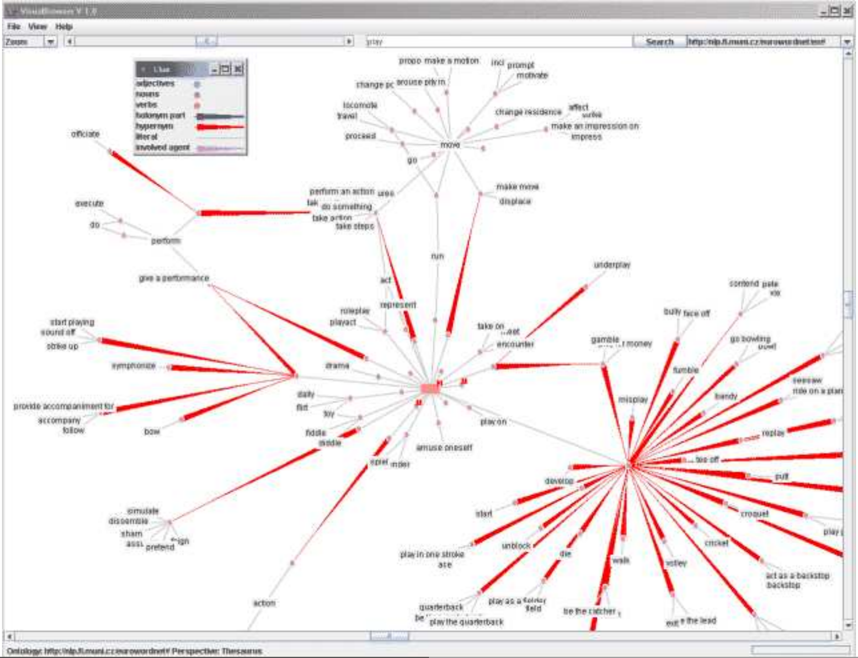
Sémantické sítě – příklad

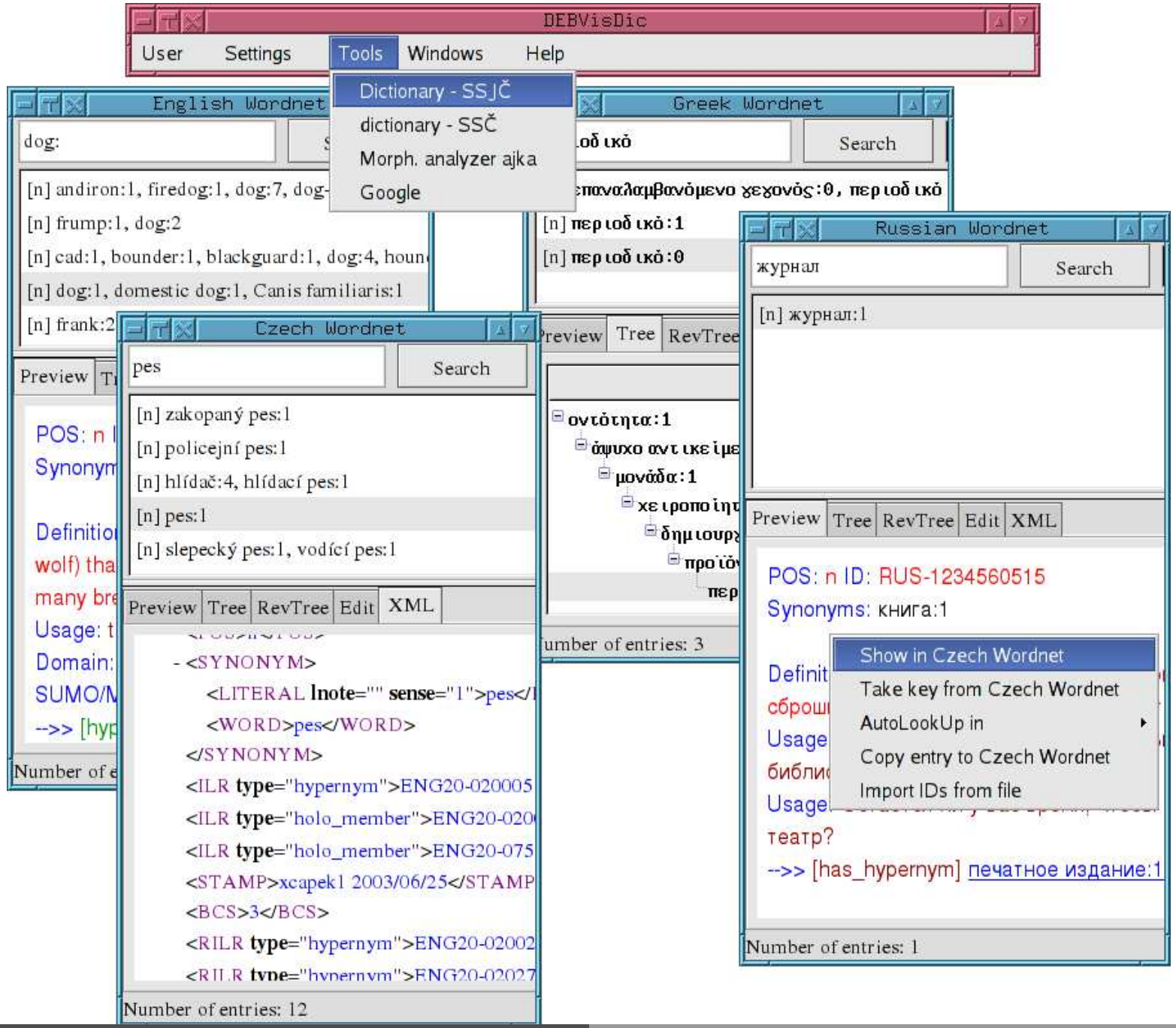


Aplikace sémantických sítí

(Princeton) WordNet – `http://wordnet.princeton.edu/`

- sématická síť 140.000 (anglických) pojmů, zachycuje:
  - synonyma, antonyma
  - hyperonyma, hyponyma
  - odvozenost a další jazykové vztahy
- jednotka synset – synonymická řada
  zachycuje slabá synonyma (*near-synonyms*)
- tvoří se národní wordnety (navázané na anglický WN)
  český wordnet – cca 30.000 pojmů
- nástroj na editaci národních wordnetů – DEBVisDic, vyvinutý na FI MU
- VisualBrowser – `http://nlp.fi.muni.cz/projekty/visualbrowser/`
  nástroj na vizualizaci (sémantických) sítí, vznikl jako DP na FI MU


File View Help
Search
officiate
execute
do
perform
give a performance
start playing
sound off
strike up
symphonize
provide accompaniment for
accompany
follow
bow
simulate
dissemble
sham
pretend
action
move
go
proceed
travel
change residence
make an impression on
impress
make move
displace
perform an action
do something
take action
take steps
run
act
represent
roleplay
playact
drama
underplay
take on
encounter
gamble
misplay
fumble
bully
go bowling
bandy
replay
tee off
putt
croquet
cricket
act as a backstop
backstop
volley
walk
die
unblock
start
develop
play in one stroke
ace
quarterback
play the quarterback
be the catcher
play on
amuse oneself
fiddle
diddle
daily
flirt
toy
motivate
prompt
affect
Ontology: http://nlp.fi.muni.cz/eurowordnet/ Perspective: Thesaurus


DEBVisDic
User
Settings
Tools
Windows
Help
Dictionary - SSJČ
dictionary - SSČ
Morph. analyzer ajka
Google
English Wordnet
dog:
[n] andiron:1, firedog:1, dog:7, dog-
[n] frump:1, dog:2
[n] cad:1, bounder:1, blackguard:1, dog:4, houn
[n] dog:1, domestic dog:1, Canis familiaris:1
[n] frank:2
Preview
POS: n
Synonym
Definitio
wolf) tha
many bre
Usage: t
Domain:
SUMO/M
-->> [hyp
Number of e
Greek Wordnet
Search
[n] περιοδικό:1
[n] περιοδικό:0
Preview
Tree
RevTree
Number of entries: 3
Russian Wordnet
журнал
Search
[n] журнал:1
Preview
Tree
RevTree
Edit
XML
POS: n ID: RUS-1234560515
Synonyms: книга:1
Show in Czech Wordnet
Take key from Czech Wordnet
AutoLookUp in
Copy entry to Czech Wordnet
Import IDs from file
театр?
-->> [has_hypernym] печатное издание:1
Number of entries: 1
Czech Wordnet
pes
Search
[n] zakopaný pes:1
[n] policejní pes:1
[n] hlídač:4, hlídací pes:1
[n] pes:1
[n] slepecký pes:1, vodící pes:1
Preview
Tree
RevTree
Edit
XML
- <SYNONYM>
<LITERAL lnote="" sense="1">pes</
<WORD>pes</WORD>
</SYNONYM>
<ILR type="hypernym">ENG20-020005
<ILR type="holo_member">ENG20-020
<ILR type="holo_member">ENG20-075
<STAMP>xcapek1 2003/06/25</STAMP
<BCS>3</BCS>
<RILR type="hypernym">ENG20-02002
Number of entries: 12

Slovníky a specializované lexikony

Slovníky typicky obsahují:

- specifikace formy:
  - grafická podoba – alternativy, dělení, velká počáteční písmena
  - zvuková podoba – výslovnost a její alternativy, slabiky, přízvuk, výška
- gramatické (morfo-syntaktické) informace – slovní druh a příslušné gramatické kategorie, morfologický vzor?
- specifikace významu – hierarchie

slovník uvádí významy listémů, encyklopedie informace o jejich denotátech

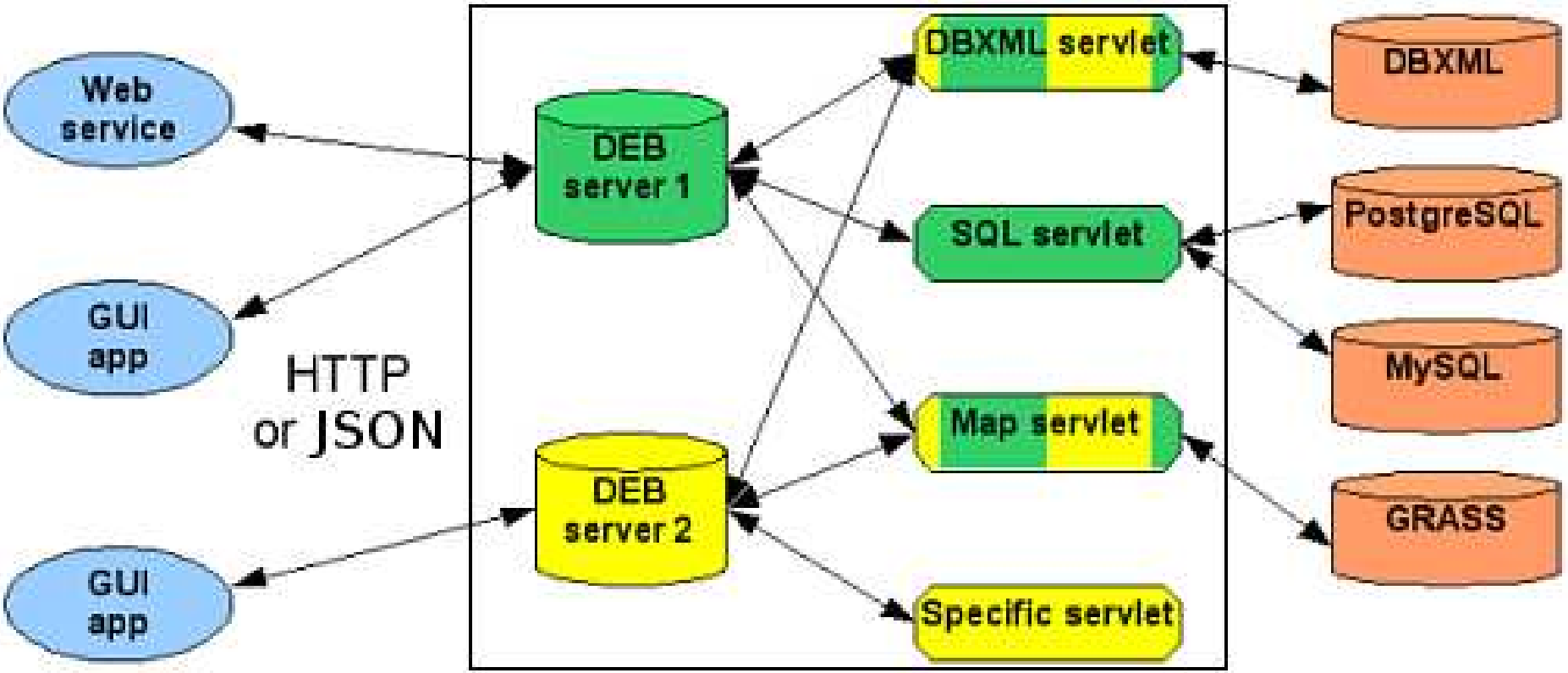
Slovník spisovné češtiny: tetřev, -a m velký lesní pták z příbuzenstva kura domácího [x] *tokat jako tetřev*, expr. být slepě zamilován; tetřeví příd. *tetřeví tokání*, *tetřeví slepice*.

Encyklopedie Diderot: tetřev, Tetrao, rod hrabavých ptáků, kteří obývají pásmo jehličnatých lesů severní polokoule. V ČR žije dnes již vzácně tetřev hlušec (Tetrao urogallus). Největší z lesních kurů, kohout dosahuje hmotnosti až 6 kg.

specializované lexikony a encyklopedie (znalost odborníků a rozdílné předpoklady a pohledy)

DEB – platforma pro vývoj slovníků

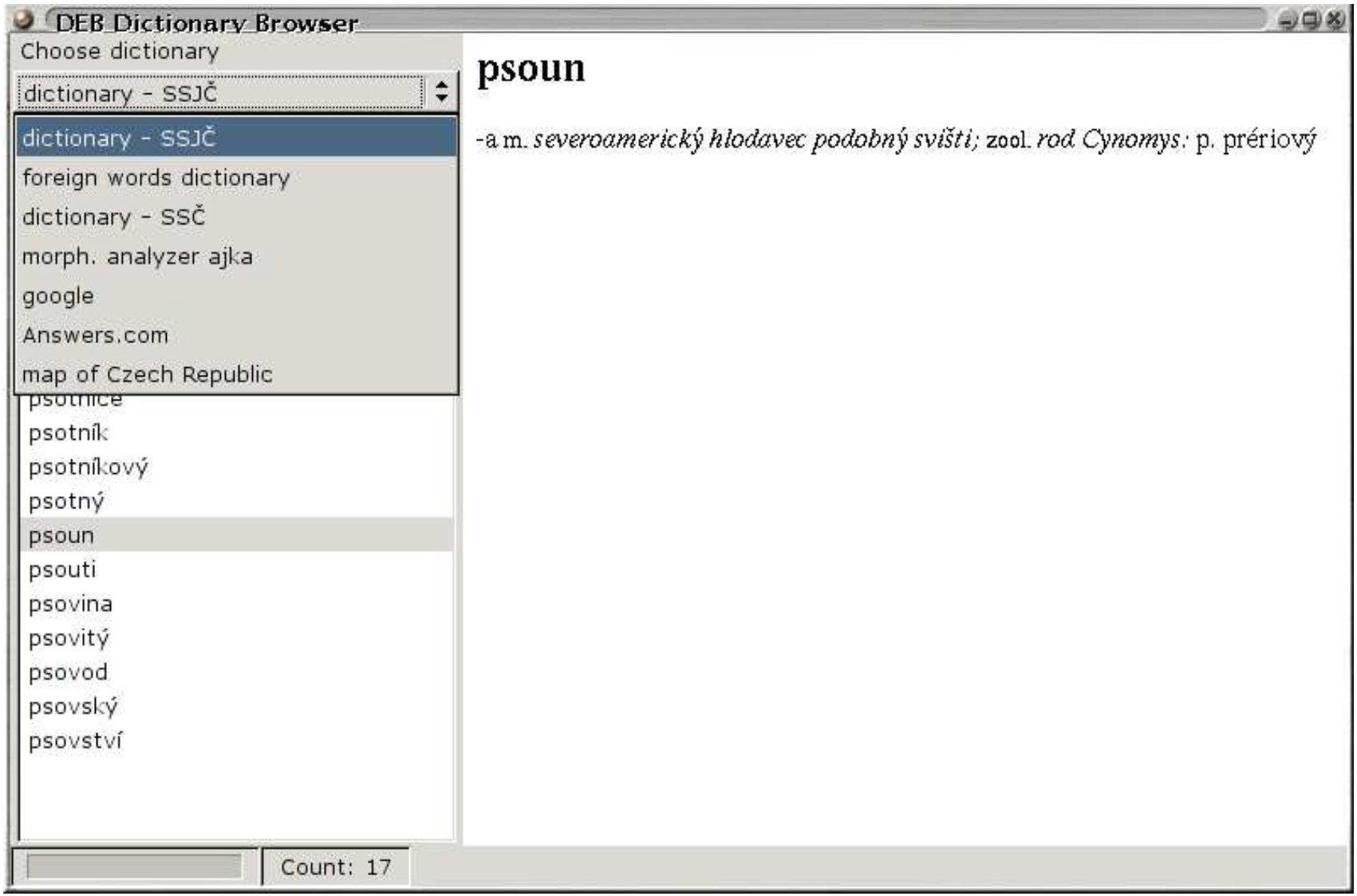
- Dictionary Editor and Browser, DEB
- platforma pro vývoj systémů na psaní slovníků (*dictionary writing systems*, DWS)
  - `http://deb.fi.muni.cz/`
  - pracuje s hesly ve formě XML struktury
- striktní klient-server architektura
- server
  - specializované moduly – *servlety*
  - databázové úložiště
- klient
  - jen jednoduchá funkcionalita
  - GUI i web rozhraní – postavený na *Mozilla Engine*



DEB používá komunikaci typu AJAX

DEBDict – příklad DEB klienta

- přehledné prohledávání slovníků s různou strukturou
- původně určený pro demo základních funkcí
- dostupný jako instalovatelné rozšíření Firefoxu i jako vzdálená webová služba
- vícejazyčné uživatelské rozhraní (angličtina, čeština, další lze snadno doplnit)
- dotazy do několika XML slovníků s různou strukturou, výsledky jsou zpracovány XSLT transformací
- autentizace – uživatelé mají různá práva přístupu ke slovníkům
- napojení na externí služby:
  - český morfologický analyzátor
  - externí webové služby (Google, Answers.com, Wikipedia)
  - geografický informační systém – zobrazení geografických odkazů přímo na mapě


DEB Dictionary Browser
Choose dictionary
dictionary - SSJČ
dictionary - SSJČ
foreign words dictionary
dictionary - SSČ
morph. analyzer ajka
google
Answers.com
map of Czech Republic
psotnice
psotník
psotníkový
psotný
psoun
psouti
psovina
psovitý
psovod
psovský
psovství
psoun
-a m. severoamerický hlodavec podobný svišti; zool. rod Cynomys: p. prériový
Count: 17

DEB – platforma pro vývoj slovníků

- další aplikace:
  - DEBVisDic – editor wordnetů
  - Cornetto – editor lexikální databáze (University of Amsterdam)
  - TeDi – terminologický slovník
  - FaNUK – slovník anglických příjmení (University of West England, Oxford University Press)
  - ...
- použitá v 22 mezinárodních projektech
- DEB server v Brně využívá více než 1600 registrovaných uživatelů

České valenční lexikony

specializované lexikony slovesných valencí:

- syntaktické valenční rámce Brief (FI MU, od 1997) cca 15,000 sloves:

```
lámat <v>hPTc4,hPTc4-hTc7,hPc3-hTc4
```

- valenční rámce v českém wordnetu (FI MU 2000), cca 3,000 slovesných literálů (sloveso+význam):

```
synset: lámat:3, dobývat:1, těžit:2
valence: kdo1*AG(person:1)=co4*SUBS(substance:1)
valence: co1*AG(institution:1)=co4*SUBS(substance:1)
```

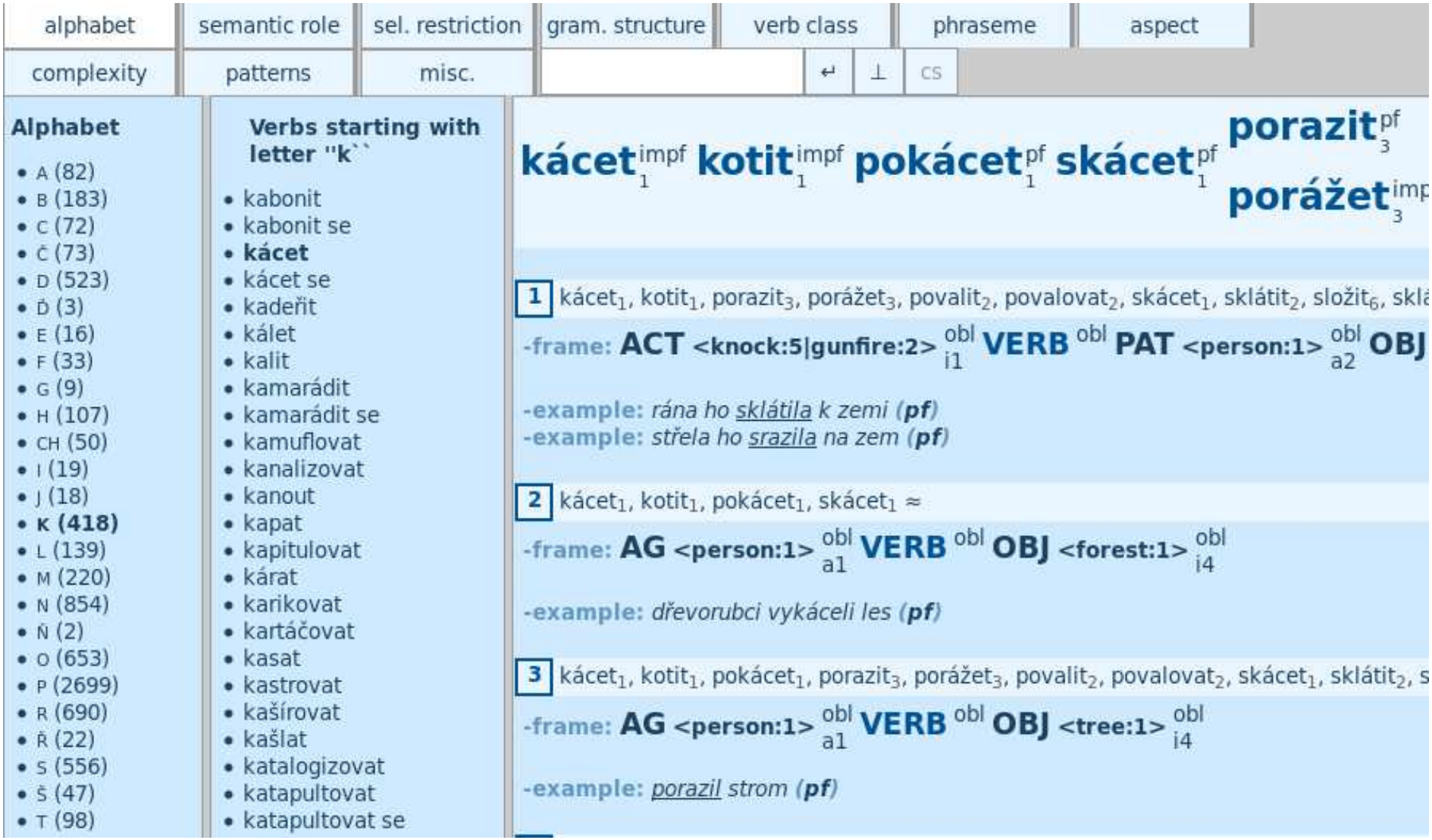
- pražský lexikon Vallex 1.0, na začátku roku 2005 cca 1,000 sloves (teď snad až 4,000):

```
~ impf: lámat
+ ACT(1;obl) PAT(4;obl)
```

Valeční lexikon VerbaLex

- vznikl na začátku roku 2005, využívá všech dostupných zdrojů
- edituje se ve formulářovém editoru nebo v jednoduchém textovém formátu, který se pro další zpracování převádí do XML
- vlastnosti:
  - dvouúrovňové sémantické role
  - odkazy na hypero/hyponymickou hierarchii v českém wordnetu
  - odlišení životnosti a neživotnosti větných členů
  - implicitní pozice slovesa
  - valenční rámce se odkazují na číslované významy sloves
- exporty z XML do HTML pro prohlížení a PDF pro tisk

VerbaLex v HTML



Využití valencí v sémantické analýze

reprezentace slovesného rámce:

1. syntaktické rysy:

   dávat něco$_{\text{neživ.NP, 4.pád, bez předložky}}$
   někomu$_{\text{živ.NP, 3.pád, bez předložky}}$

2. sémantické rysy:

   dávat Patiens Addressee

3. funkce významu:

   dávat $x\ y \ldots (o(o\pi)(o\pi))_\omega$, slovesný objekt
   $dávat/(o(o\pi)(o\pi))_\omega \iota\iota \qquad x \ldots \iota \qquad y \ldots \iota : s_{wt}y, s \ldots (o\iota)_{\tau\omega}$

překlad z valenčního výrazu do funkce významu:

$$\text{typ argumentu} = \text{typ} \begin{cases} \blacktriangleright \text{ jmenné skupiny} \\ \blacktriangleright \text{ příslovečné fráze} \\ \blacktriangleright \text{ vedlejší věty} \\ \blacktriangleright \text{ infinitivu} \end{cases}$$

Problémy sémantiky s jazykovými zdroji

Problémy jazykových zdrojů:

- nejsou dostupné pro každý jazyk
- neobsahují všechna slova
- neobsahují dost kombinací slov, frází
- neobsahují všechny významy
- neobsahují všechny relace
- naopak obsahují i (velmi) málo frekventované významy/relace (jak – spojka/zvíře, s – předložka/citoslovce, tři – číslovka/sloveso)
- relace nejsou stejně strukturované pro různé slovní druhy (H/H relace moc nefunguje pro přídavná jména, slovesa)

Distribuční sémantické modely

alternativa – automatické distribuční sémantické modely

- také vektorové modely (*vector-space models*)
- slova/fráze/dokumenty nahrazujeme
  body v N-rozměrném vektorovém prostoru (vektory)
  (kde N může být velké číslo – stovky tisíc)
- modely se počítají automaticky z rozsáhlých textových sad
- dosahují vyšší pokrytí, ale menší přesnost než "ruční" jazykové zdroje
- primární počítaná sémantická operace – podobnost

Podobnost dokumentů a slov

Podobnost dokumentů:

- důležitá např. pro vyhledávání informací
- dokument (dotaz) = vektor frekvencí (TF-IDF frekvenčních skóre) slov

doc1: Hotel byl krásný, ale personál hotelu nepříjemný.
doc2: Hotel je standardní a jídlo v hotelu vynikající.
query: hotel a jídlo

	doc1	doc2	query
~~a~~	~~0~~	~~1~~	~~1~~
~~ale~~	~~1~~	~~0~~	~~0~~
~~byl~~/být	1	2	0
hotel	2	2	1
~~hotelu~~/hotel	~~1~~	~~1~~	~~0~~
~~je~~/být	~~0~~	~~1~~	~~0~~
jídlo	0	1	1
krásný	1	0	0
nepříjemný	1	0	0
personál	1	0	0
standardní	0	1	0
~~v~~	~~0~~	~~1~~	~~0~~
vynikající	0	1	0

$$vec_{doc1} = \langle 1, 2, 0, 1, 1, 1, 0, 0 \rangle$$
$$vec_{doc2} = \langle 2, 2, 1, 0, 0, 0, 1, 1 \rangle$$
$$vec_{query} = \underbrace{\langle 0, 1, 1, 0, 0, 0, 0, 0 \rangle}_{8}$$

snížení (prokletí) dimensionality:

- výběr rysů (*feature selection*) – stop slova, frekventovaná slova, ...
- extrakce rysů (*feature extraction*) – lemmatizace/stemming, latentní sémantická analýza, ...

Podobnost dokumentů

2 dokumenty jsou podobné ⇔ jsou podobné jejich vektory

podobnost vektorů se určuje cosinovou podobností:

$$\frac{\vec{a} \cdot \vec{b}}{||\vec{a}|| \cdot ||\vec{b}||}$$

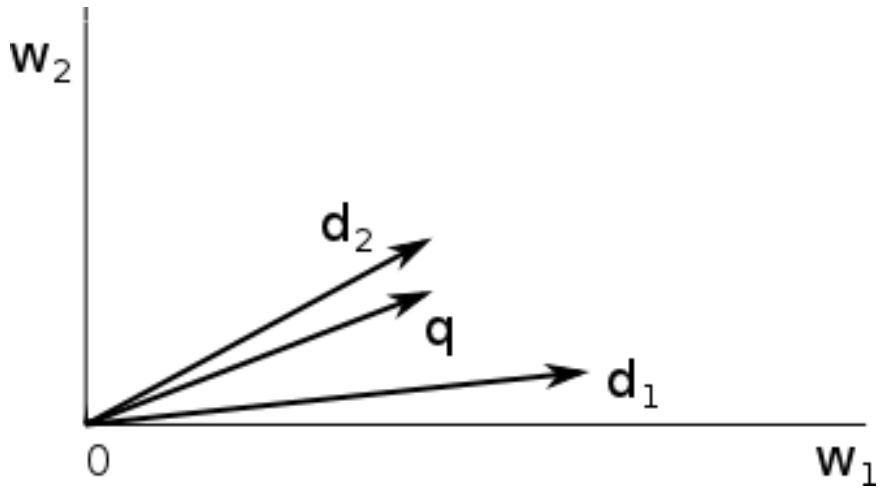
	doc1	doc2	query
být	1	2	0
hotel	2	2	1
jídlo	0	1	1
krásný	1	0	0
nepříjemný	1	0	0
personál	1	0	0
standardní	0	1	0
vynikající	0	1	0

doc1: Hotel byl krásný, ale personál hotelu nepříjemný.
doc2: Hotel je standardní a jídlo v hotelu vynikající.
query: hotel a jídlo

(normalizovaný skalární součin vektorů, cosinus úhlu mezi vektory)

$sim_{cos}(doc_1, query) = 0.5$

$sim_{cos}(doc_2, query) = 0.64$



Podobnost slov

analogicky slovo = vektor frekvencí slova v dokumentech

	doc1	doc2	query
být	1	2	0
hotel	2	1	1
jídlo	0	1	1
krásný	1	0	0
nepříjemný	1	0	0
personál	1	0	0
standardní	0	1	0
vynikající	0	1	0

$$vec_{\text{standardní}} = \langle 0, 1, 0 \rangle$$
$$vec_{\text{vynikající}} = \langle 0, 1, 0 \rangle$$

2 slova jsou podobná $\Leftrightarrow$ jsou podobné jejich vektory

(to samozřejmě funguje lépe na velkých datech)

Reprezentace slov

reálně se místo dokumentů používají kontexty

... jsou na látky obsažené v čokoládě (kofein, theobromin) mimořádně citliví a nedokáží je ...
... kofein, který najdete v čokoládě, a theobromin působí stimulačně na centrální nervový ...
... se skrývá mimo jiné fenyletylamin a theobromin , přičemž mu jsou přisuzovány opojné ...
... podoba v čaji se nazývá theofylin a v kakau theobromin – účinky jsou prakticky stejné ...
... celospolečensky tolerované drogy, jako kofein, theobromin , nebo nikotin ...

z kontextů poznáme (odhadneme, kontexty určují) význam slova

(theobromin – *látka vyskytující se v čokoládě s podobným stimulačním účinkem jako kofein*)

Reprezentace slov

místo frekvencí slov –
skóre vzájemné informace (*Mutual Information (MI) score*)

MI skóre pro pravděpodobnostní jevy – *vyskytují se jevy X a Y spolu více, než kdyby byly nezávislé?*

$$MI(X, Y) = log_2 \frac{P(X, Y)}{P(X)P(Y)}$$

MI skóre pro slova/kontexty – *vyskytuje se slovo word v kontextu context více, než kdyby byly nezávislé?*

$$MI(word, context) = log_2 \frac{P(word, context)}{P(word)P(context)}$$

může se upravovat vážením (*weighting*) a vyhlazováním (*smoothing*)

Zapouzdření slov (Word Embedding)

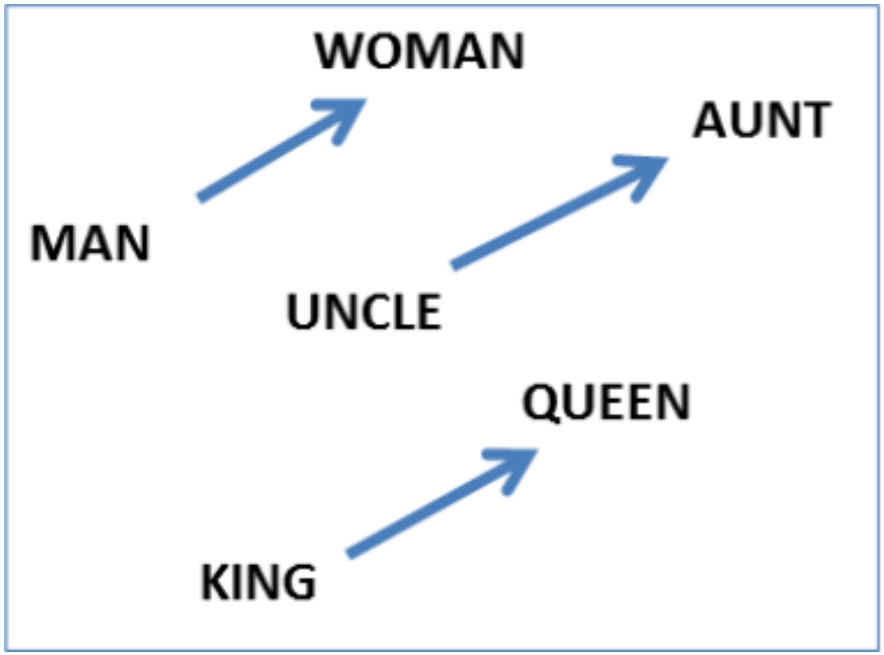
- jiný způsob reprezentace významu slov ve vektorovém prostoru
- na principu extrakce rysů – počet rysů stanovíme (třeba 1000)
- slovo inicializujeme jako náhodný vektor v prostoru rysů
- cyklicky upravujeme vektory tak, abychom maximalizovali podmíněnou pravděpodobnost mezi slovem a jeho kontexty

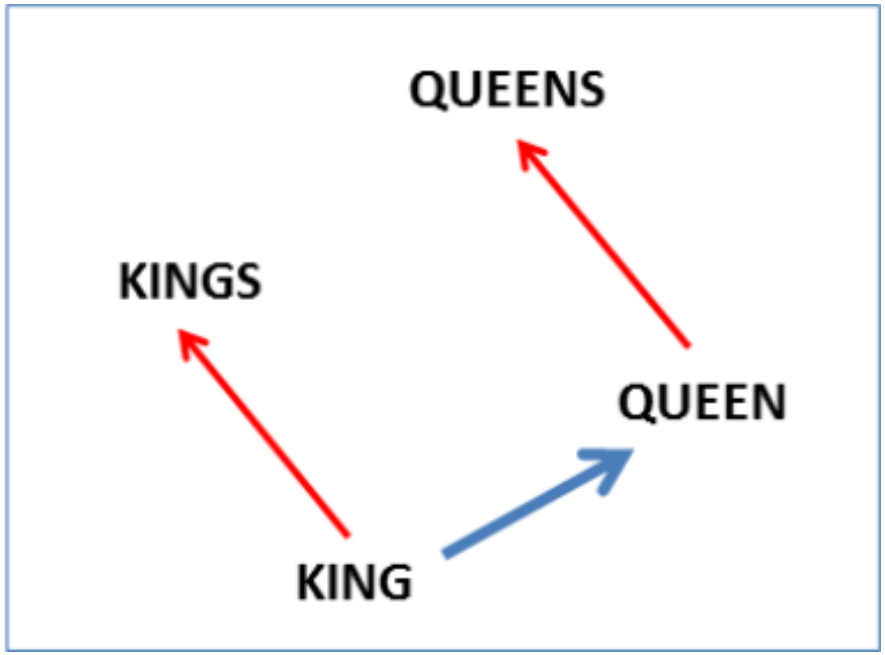
$$\arg\max_{\theta} \prod_{(w,c)\in D} P(c|w;\theta)$$

- algoritmy – word2vec (Mikolov, Google, princip učení neuronové sítě), GloVe (Pennington et al, Stanford Uni, faktorizace matic)
  pro kvalitní výstupy je potřeba velmi velká data (miliardy slov)
  existují rozšíření na fráze (phrase2vec) a dokumenty (doc2vec)

Zapouzdření slov (Word Embedding)

sémantické vlastnosti výsledných vektorů





(příklady od T. Mikolova)

Zapouzdření slov (Word Embedding)

sémantické vlastnosti výsledných vektorů

operace s vektory	nejbližší výsledný vektor
Paris - France + Italy	Rome
bigger - big + cold	colder
sushi - Japan + Germany	bratwurst
Cu - copper + gold	Au
Windows - Microsoft + Google	Android
Montreal Canadiens - Montreal + Toronto	Toronto Maple Leafs

(příklady od T. Mikolova)

Zapouzdření slov (Word Embedding)

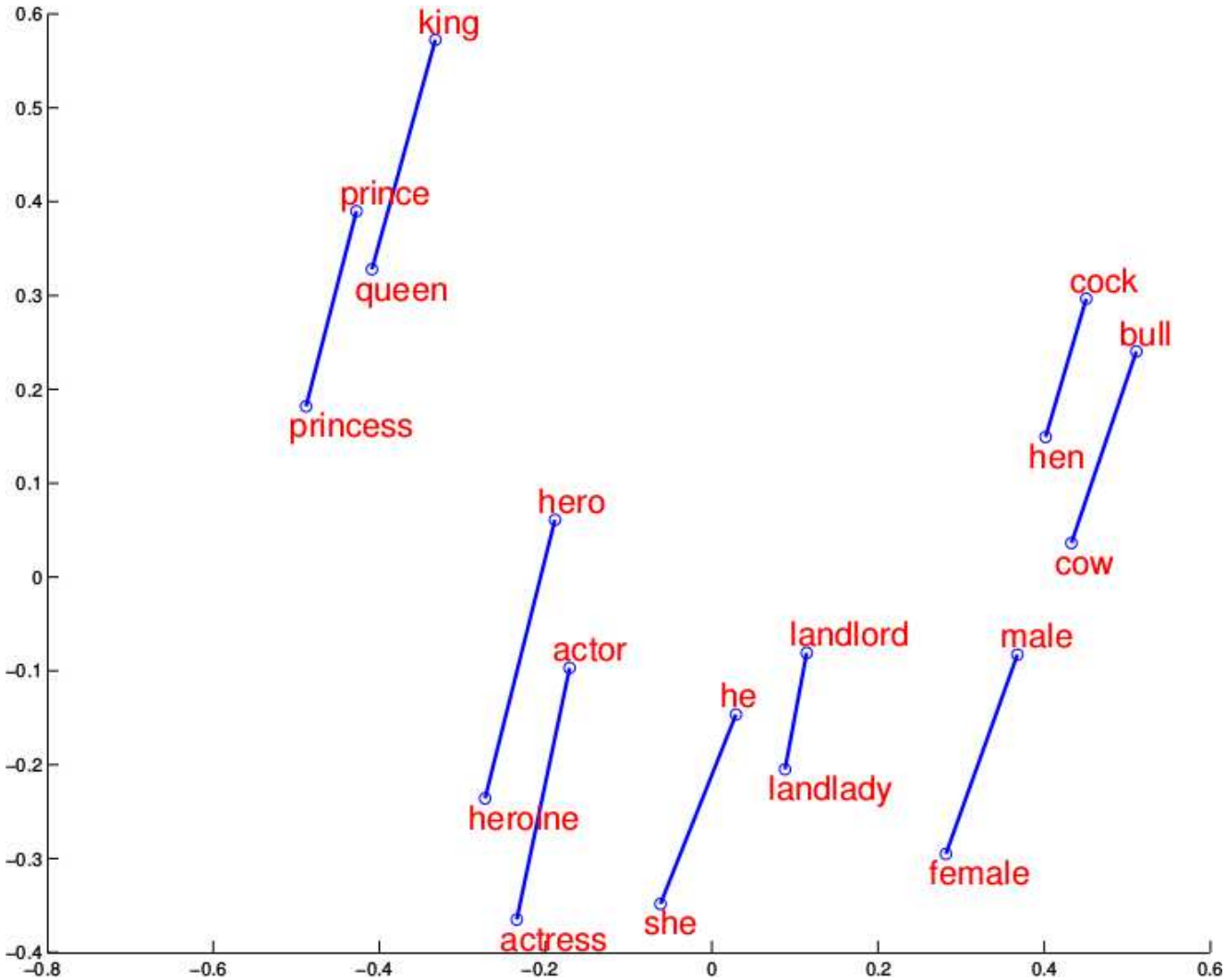
sémantické vlastnosti výsledných vektorů

operace s vektory	nejbližší vektory
Czech + currency	koruna, Czech crown, Polish zloty, CTK
Vietnam + capital	Hanoi, Ho Chi Minh City, Viet Nam, Vietnamese
German + airlines	airline Lufthansa, carrier Lufthansa
Russian + river	Moscow, Volga River, upriver, Russia
French + actress	Juliette Binoche, Vanessa Paradis

(příklady od T. Mikolova)

Zapouzdření slov (Word Embedding)

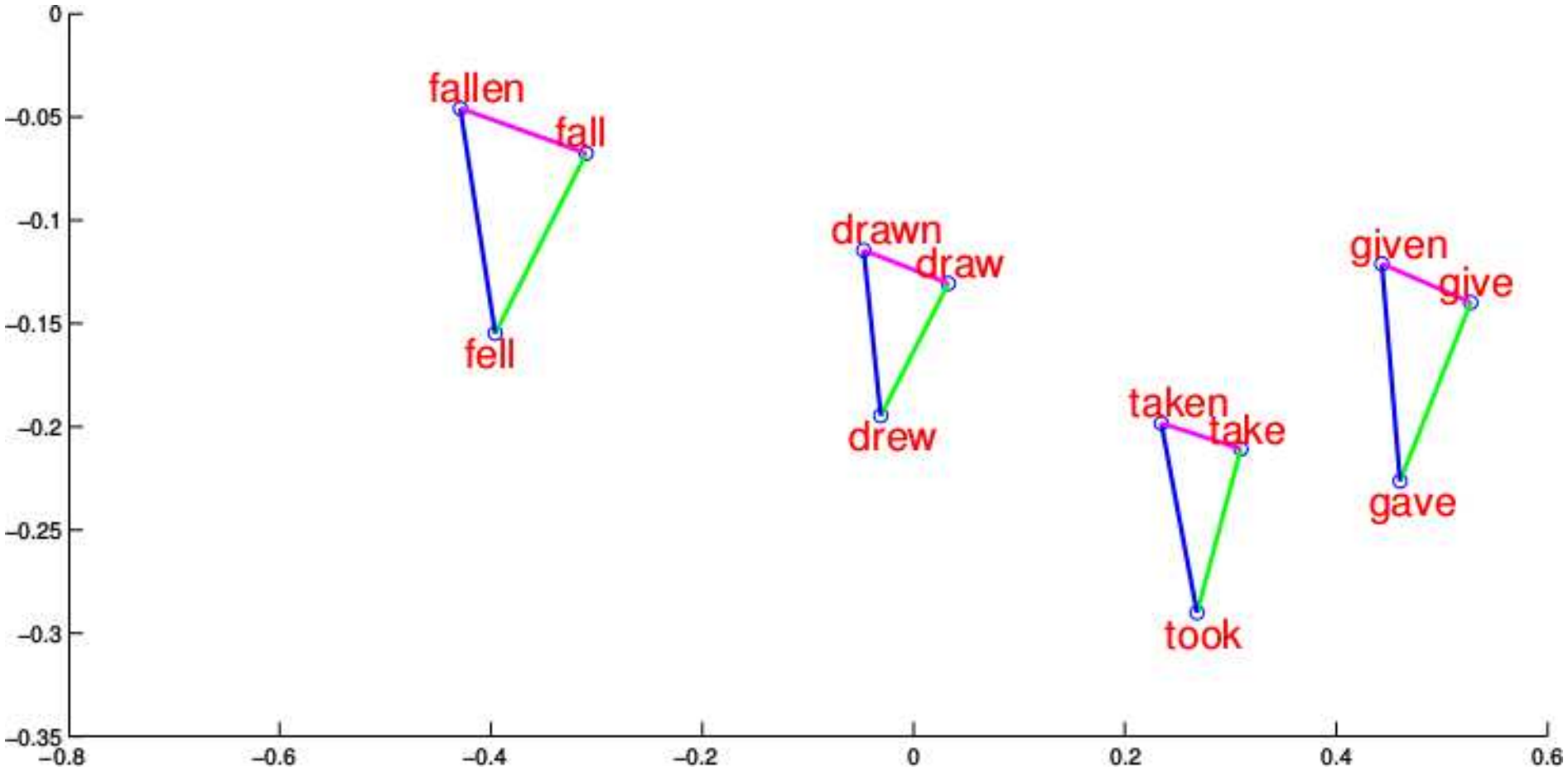
vizualizace pravidelností výsledných vektorů



(příklady od T. Mikolova)

Zapouzdření slov (Word Embedding)

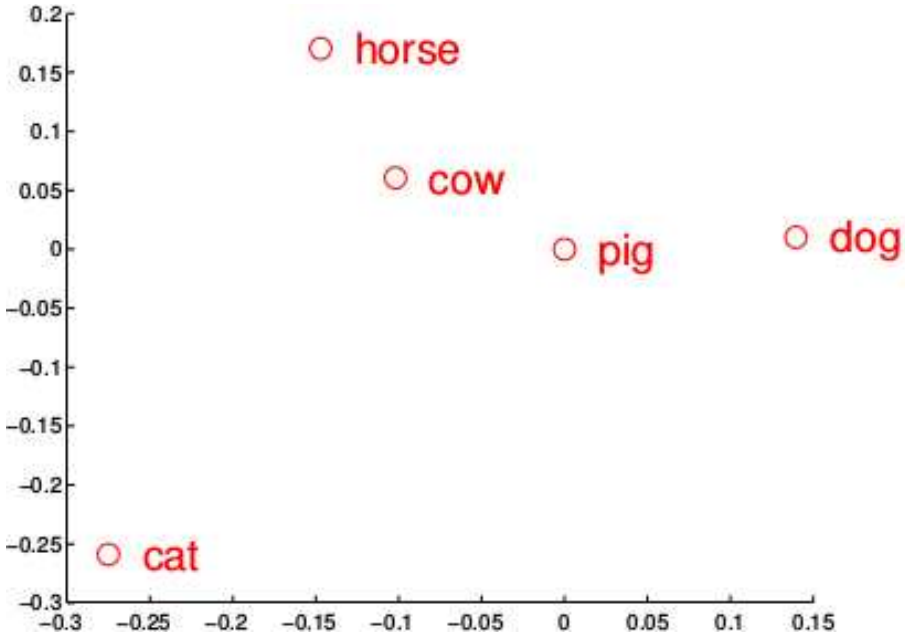
vizualizace pravidelností výsledných vektorů

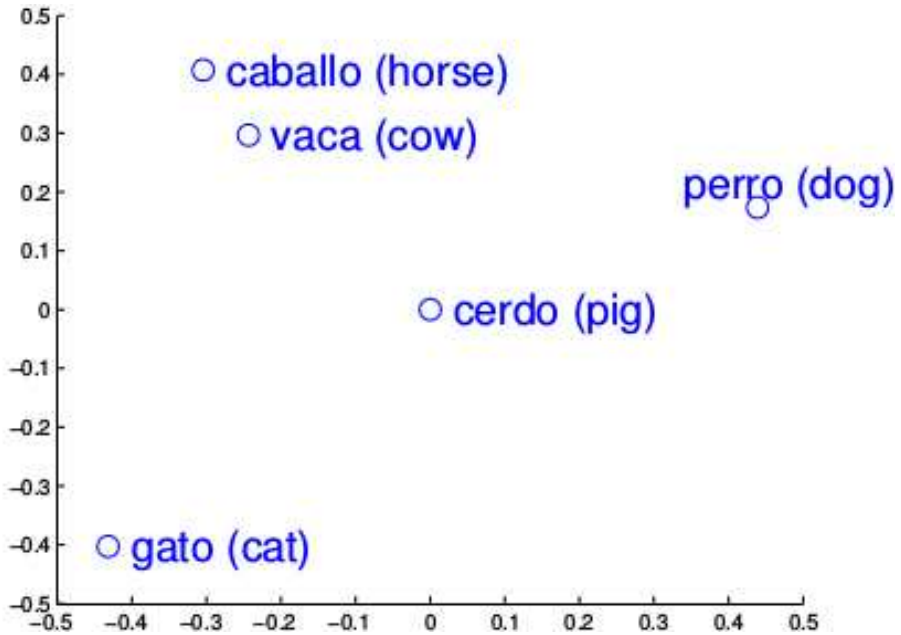


(příklady od T. Mikolova)

Zapouzdření slov (Word Embedding)

využití vektorových reprezentací pro strojový překlad
prostory různých jazyků je nutné lineárně transformovat (otočit, zmenšit)





(příklady od T. Mikolova)