

# Výpočetní sémantika a základní sémantické struktury

Aleš Horák

E-mail: [hales@fi.muni.cz](mailto:hales@fi.muni.cz)

[http://nlp.fi.muni.cz/poc\\_lingv/](http://nlp.fi.muni.cz/poc_lingv/)

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

| freq | 6.0 | 4.0 | 2.0 | 0   | -2.0 | -4.0 | -6.0 | tanec |
|------|-----|-----|-----|-----|------|------|------|-------|
| 6.0  | 6   | 0   | 4.2 | –   |      |      |      |       |
| 4.0  | 5   | 0   | 1.9 | –   |      |      |      |       |
| 2.0  | 6   | 0   | 0.8 | –   |      |      |      |       |
| 0    | 0   | 6   | –   | 0.2 |      |      |      |       |
| -2.0 | 0   | 18  | –   | 1.3 |      |      |      |       |
| -4.0 | 0   | 7   | –   | 1.3 |      |      |      |       |
| -6.0 | 0   | 18  | –   | 1.3 |      |      |      |       |

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

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

| a_modifier       | 3,218      | 5,764      | 1.70 | 2.20 |
|------------------|------------|------------|------|------|
| modernizovaný    | <u>65</u>  | 0          | 9.0  | –    |
| sovětský         | <u>318</u> | 0          | 8.4  | –    |
| vyprošťovací     | <u>29</u>  | 0          | 8.0  | –    |
| zničený          | <u>38</u>  | 0          | 7.4  | –    |
| zastaralý        | <u>24</u>  | 0          | 6.9  | –    |
| mostní           | <u>15</u>  | 0          | 6.9  | –    |
| Wittmannův       | <u>11</u>  | 0          | 6.8  | –    |
| lehký            | <u>103</u> | <u>5</u>   | 6.9  | 2.4  |
| moderní          | <u>40</u>  | <u>238</u> | 4.5  | 7.0  |
| latinskoamerický | <u>8</u>   | <u>90</u>  | 5.8  | 8.7  |
| povinný          | <u>7</u>   | <u>91</u>  | 3.3  | 6.9  |
| originální       | 0          | <u>69</u>  | –    | 6.8  |
| společenský      | 0          | <u>144</u> | –    | 6.8  |
| lidový           | 0          | <u>157</u> | –    | 7.0  |
| dvořákov         | 0          | <u>38</u>  | –    | 7.2  |
| irský            | 0          | <u>69</u>  | –    | 7.3  |
| country          | 0          | <u>77</u>  | –    | 7.8  |
| výrazový         | 0          | <u>50</u>  | –    | 7.8  |
| scénický         | 0          | <u>63</u>  | –    | 8.0  |
| dobový           | 0          | <u>104</u> | –    | 8.1  |
| rituální         | 0          | <u>62</u>  | –    | 8.2  |
| slovanský        | 0          | <u>104</u> | –    | 8.2  |
| hříšný           | 0          | <u>92</u>  | –    | 8.7  |
| bříšni           | 0          | <u>329</u> | –    | 10.3 |
| orientální       | 0          | <u>404</u> | –    | 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    | <u>5</u> | 0        | 1.5  | --   |
| obrázek  | <u>5</u> | 0        | 1.4  | --   |
| obraz    | <u>6</u> | 0        | 1.4  | --   |
| noc      | <u>4</u> | 0        | 1.3  | --   |
| barva    | <u>4</u> | <u>9</u> | 0.9  | 2.1  |

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

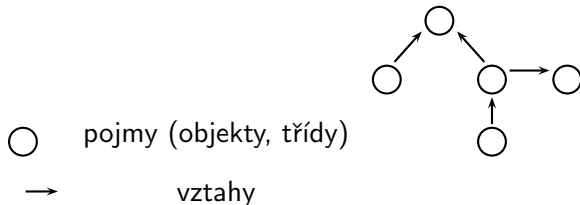
- ▶ slova **nadřazená**/**podřazená** (*superordinate/subordinate*)
- ▶ z logického pohledu  $u_j$  je **hyponymum**  $w_i \Leftrightarrow$ 
  - **extenzionálně** –  $class(u_j) \subset class(w_i)$
  - **vyplývání** –  $property(x, u_j) \Rightarrow property(x, w_i)$
- ▶ hypero/hyponymie je obvykle **tranzitivní**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

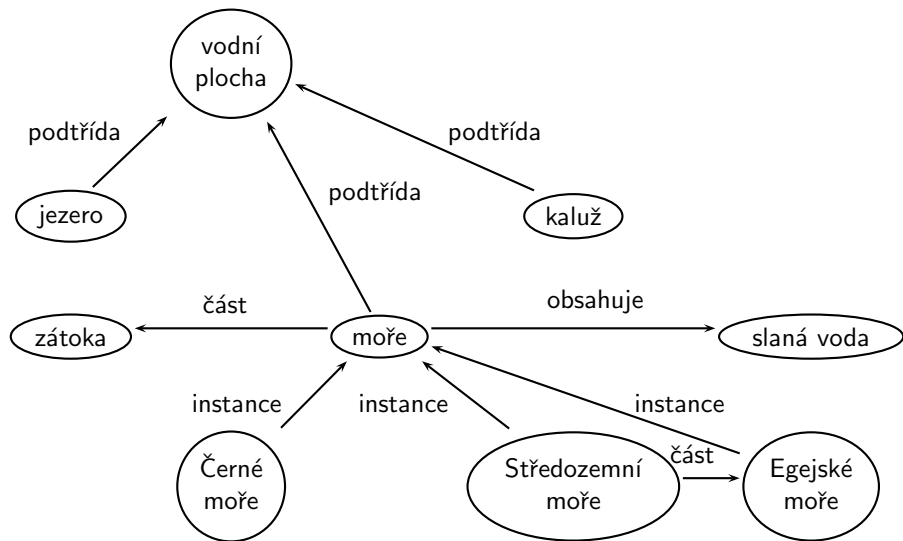


- ▶ nejdůležitější vztahy:

- **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/>

- ▶ sémantická 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



The screenshot displays the DEBVisDic application with several wordnet windows open:

- English Wordnet:** Search for "dog:". Results include: [n] andiron:1, fire-dog:1, dog:7, dog; [n] frump:1, dog:2; [n] cad:1, boulder:1, black-guard:1, dog:4, hound:1; [n] dog:1, domestic dog:1, Canis familiaris:1; [n] frank:2.
- Greek Wordnet:** Search for "οδ ικό". Results include: [n] περ τοδ ικό:1; [n] περ τοδ ικό:θ.
- Czech Wordnet:** Search for "pes". Results include: [n] zakopaný pes:1; [n] policejní pes:1; [n] hlídač:4, hlídač pes:1; [n] pes:1; [n] slepecký pes:1, vodící pes:1.
- Russian Wordnet:** Search for "журнал". Results include: [n] журнал:1.

A menu is open over the English Wordnet window with the following options: Dictionary - SSČ, dictionary - SSČ, Morph. analyzer ajka, Google.

Another menu is open over the Russian Wordnet window with the following options: Show in Czech Wordnet, Take key from Czech Wordnet, AutoLookUp in, Copy entry to Czech Wordnet, Import IDs from file.

The Czech Wordnet window shows a detailed view of the entry "pes" with the following XML-like structure:

```

<SYNONYM>
<LITERAL Inote="" sense="1">pes</LITERAL>
<WORD>pes</WORD>
<SYNONYM>
<ILR type="hypernym">ENG20-020005
<ILR type="holo_member">ENG20-020006
<ILR type="holo_member">ENG20-075
<STAMP>xcapek1 2003/06/25</STAMP>
<BCS>3</BCS>
<RILR type="hypernym">ENG20-02002
<RILR tvne="hynernym">FNG20-02027

```

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*.

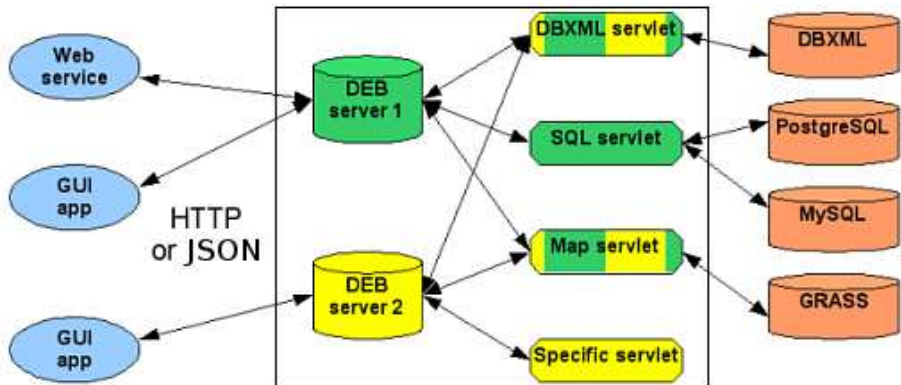
Encyklopedie Diderot: **tetřev**, Tetrao, rod hrabavých ptáků, kteří obývají pásmo jehličnatých lesů severní polokoule. V ČR žije dnes již vzácně tetřev hlušec (Tetrao urogallus). Největší z lesních kurů, kohout dosahuje hmotnosti až 6 kg.

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*

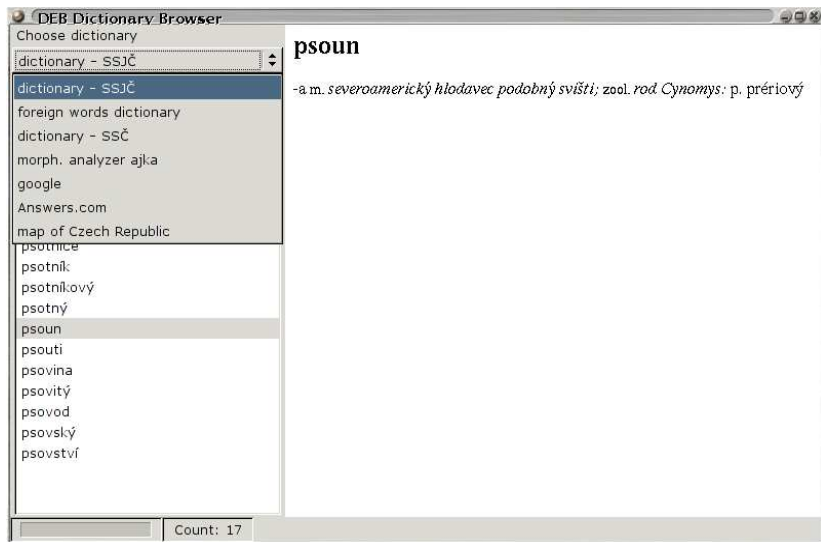




DEB používá komunikaci typu AJAX

## 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 – 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)
  - ...
- ▶ použita 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:

lázat <v>hPTc4,hPTc4-hTc7,hPc3-hTc4

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

synset: lázat:3, dobývat:1, těžít: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: lázat

+ 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

| alphabet  | semantic role   | sel. restriction   | gram. structure | verb class | phraseme | aspect |
|---|---|--|-----------------|------------|----------|--------|
| complexity  | patterns  | misc.  |                 | ↔          | ⊥        | CS     |
| <b>Alphabet</b> <ul style="list-style-type: none"> <li>• A (82)</li> <li>• B (183)</li> <li>• C (72)</li> <li>• Č (73)</li> <li>• D (523)</li> <li>• Ď (3)</li> <li>• E (16)</li> <li>• F (33)</li> <li>• G (9)</li> <li>• H (107)</li> <li>• CH (50)</li> <li>• I (19)</li> <li>• J (18)</li> <li>• <b>K (418)</b></li> <li>• L (139)</li> <li>• M (220)</li> <li>• N (854)</li> <li>• Ň (2)</li> <li>• O (653)</li> <li>• P (2699)</li> <li>• R (690)</li> <li>• Ř (22)</li> <li>• S (556)</li> <li>• Š (47)</li> <li>• T (98)</li> </ul> | <b>Verbs starting with letter "k"</b> <ul style="list-style-type: none"> <li>• kabonit</li> <li>• kabonit se</li> <li>• <b>kácet</b></li> <li>• kácet se</li> <li>• kadeřit</li> <li>• kálet</li> <li>• kalit</li> <li>• kamarádit</li> <li>• kamarádit se</li> <li>• kamuflovat</li> <li>• kanalizovat</li> <li>• kanout</li> <li>• kapat</li> <li>• kapitulovat</li> <li>• kárat</li> <li>• karikovat</li> <li>• kartáčovat</li> <li>• kasat</li> <li>• kastrovat</li> <li>• kaširovat</li> <li>• kašlat</li> <li>• katalogizovat</li> <li>• katapultovat</li> <li>• katapultovat se</li> </ul> | <p><b>kácet</b><sub>1</sub><sup>impf</sup> <b>kotit</b><sub>1</sub><sup>impf</sup> <b>pokácet</b><sub>1</sub><sup>pf</sup> <b>skácet</b><sub>1</sub><sup>pf</sup> <b>porazit</b><sub>3</sub><sup>pf</sup><br/><b>porážet</b><sub>3</sub><sup>impf</sup></p> <p><b>1</b> kácet<sub>1</sub>, kotit<sub>1</sub>, porazit<sub>3</sub>, porážet<sub>3</sub>, povalit<sub>2</sub>, povalovat<sub>2</sub>, skácet<sub>1</sub>, sklátit<sub>2</sub>, složit<sub>6</sub>, sklá<br/> <b>-frame:</b> <b>ACT</b> &lt;knock:5 gunfire:2&gt; obl <b>VERB</b> obl <b>PAT</b> &lt;person:1&gt; obl <b>OBJ</b><br/> <b>-example:</b> rána ho <i>sklátila</i> k zemi (<b>pf</b>)<br/> <b>-example:</b> střela ho <i>srazila</i> na zem (<b>pf</b>)</p> <p><b>2</b> kácet<sub>1</sub>, kotit<sub>1</sub>, pokácet<sub>1</sub>, skácet<sub>1</sub> ≈<br/> <b>-frame:</b> <b>AG</b> &lt;person:1&gt; obl <b>VERB</b> obl <b>OBJ</b> &lt;forest:1&gt; obl<br/> <b>-example:</b> dřevorubci <i>vykáceli</i> les (<b>pf</b>)</p> <p><b>3</b> kácet<sub>1</sub>, kotit<sub>1</sub>, pokácet<sub>1</sub>, porazit<sub>3</sub>, porážet<sub>3</sub>, povalit<sub>2</sub>, povalovat<sub>2</sub>, skácet<sub>1</sub>, sklátit<sub>2</sub>, s<br/> <b>-frame:</b> <b>AG</b> &lt;person:1&gt; obl <b>VERB</b> obl <b>OBJ</b> &lt;tree:1&gt; obl<br/> <b>-example:</b> <i>porazil</i> strom (<b>pf</b>)</p> |                 |            |          |        |

# Využití valencí v sémantické analýze

reprezentace **slovesného rámce**:

## 1. syntaktické rysy:

dávat něčO<sub>neživ.NP</sub>, 4.pád, bez předložky

někomu<sub>živ.NP</sub>, 3.pád, bez předložky

## 2. sémantické rysy:

dávat Patiens Addressee

## 3. funkce významu:

**dávat**  $x y \dots (o(o\pi)(o\pi))_{\omega}$ , slovesný objekt

*dávat* /  $(o(o\pi)(o\pi))_{\omega ll} \quad x \dots l \quad y \dots l : s_{wt}y, s \dots (ol)_{T\omega}$

**překlad** z valenčního výrazu do funkce významu:

typ argumentu = typ {

- ▶ jmenné skupiny
- ▶ příslovečné fráze
- ▶ vedlejší věty
- ▶ infinitivu



# 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     |

$$\text{vec}_{doc1} = \langle 1, 2, 0, 1, 1, 1, 0, 0 \rangle$$

$$\text{vec}_{doc2} = \langle 2, 2, 1, 0, 0, 0, 1, 1 \rangle$$

$$\text{vec}_{query} = \langle 0, 1, 1, 0, 0, 0, 0, 0 \rangle$$

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**  $\Leftrightarrow$  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.

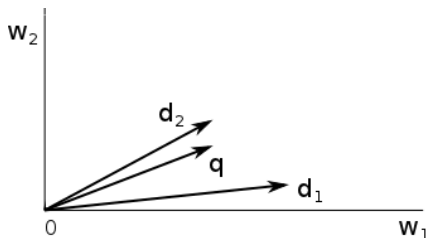
doc2: Hotel je standardní a jídlo v hotelu vynikající.

query: hotel a jídlo

(normalizovaný skalární součin vektorů,  
 cosinus úhlu mezi vektory)

$$\text{sim}_{\text{cos}}(\text{doc}_1, \text{query}) = 0.5$$

$$\text{sim}_{\text{cos}}(\text{doc}_2, \text{query}) = 0.64$$



# 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**

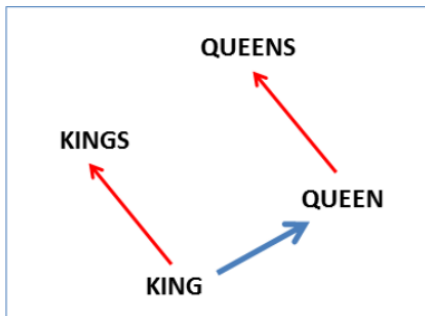
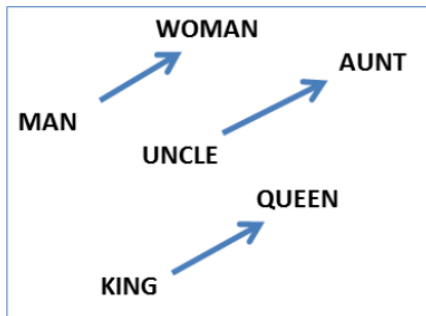
$$\arg \max_{\theta} \prod_{(w,c) \in D} P(c|w; \theta)$$

- ▶ 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)

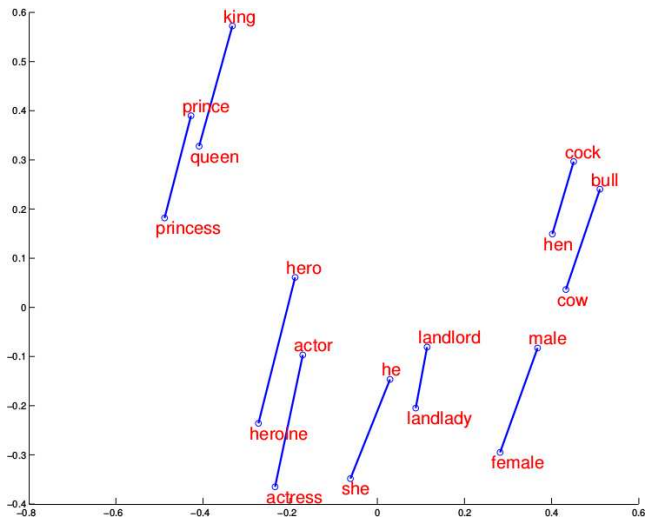
sémantické vlastnosti výsledných vektorů

| 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)

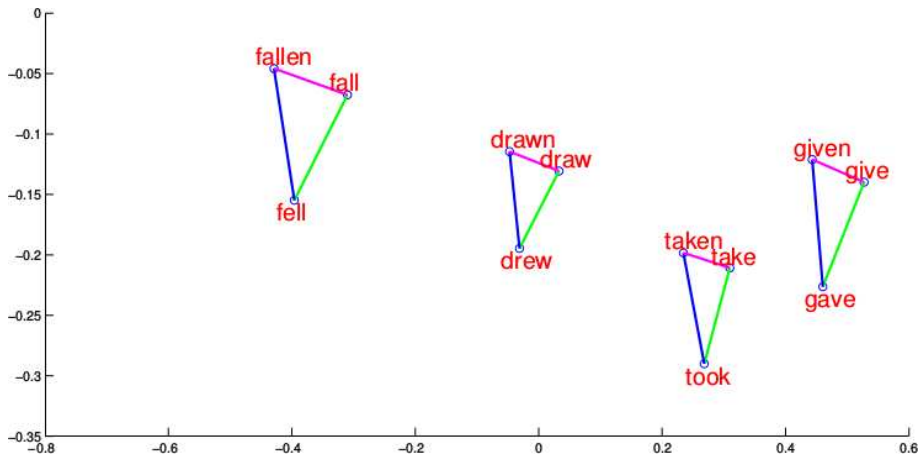
vizualizace pravidelností výsledných vektorů



(příklady od T. Mikolova)

# Zapouzdření slov (Word Embedding)

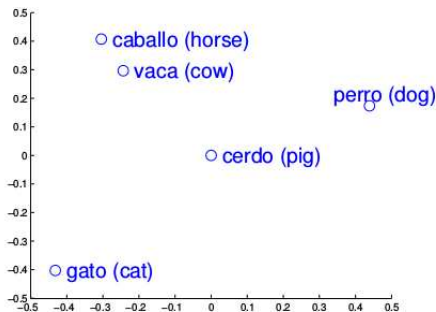
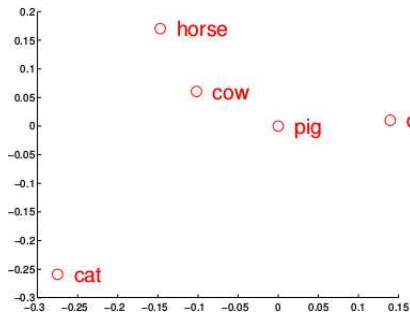
vizualizace pravidelností výsledných vektorů



(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)