Aleš Horák

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

Obsah:

- Sémantika
- Slovníky a encyklopedie
- Sémantické sítě
- Reprezentace slovesných valencí

Sémantika

studium významu – rozdílné, i když překrývající se přístupy různých vědeckých disciplín:

- **filosofie** – Jak je možné, že něco vůbec něco znamená?
  Jaký typ relace musí být mezi X a Y, aby X znamenalo Y? (filosofie jazyka)
- **psychologie** – psycholingvistika – experimentální studie, jak jsou významy reprezentovány v mysli a jaké mechanismy ovlivňují při kódování a dekódování zpráv (délka odezvy u konkrét a abstrakt se liší)
- **neurologie** – jak jsou psychologické stavy a procesy *implementovány* na úrovni neuronů

Význam v jazyce

Rozdělení studia významu v jazyce:

- **lexikální sémantika**
- **gramatická sémantika** – větné fráze, slovotvorba
- **logická sémantika** – výroková, predikátová a vyšší logiky
- **lingvistická pragmatika**

entail = znamenat, vyplývat; nutnost a očekávanost

1. X přestal zpívat ?→? X nepokračoval ve zpěvu
2. X je kočka ?→? je zvíře
3. X je v jiném stavu ?→?X je žena
4. X je fyzikální objekt ?→? X má hmotnost
5. X je čtyřnožec ?→? X má čtyři nohy
6. X je žena Y ?→? X není dcera Y

Princip kompozicionality

Význam složeného tvrzení je funkcí významu jednotlivých komponent.

(je určován, je odhadnutelný, každá složka hraje význam?)
omezení PK: idiomy, ustrnulé metafory, kolokace, klišé

listém je jazykový výraz, jehož význam není určen významy jeho částí (pokud existují), a který si tedy uživatel jazyka musí zapamatovat jako kombinaci formy a významu.

každý neodvozený **lexém** je listém.

Problémy při analýze přirozeného jazyka

- víceznačnost
- anaforické výrazy
- indexické výrazy
- nejasnost
- nekompozicionalita
- struktura promluvy
- metonymie
- metafory

Víceznačnost

- *ambiguity*
- **víceznačnost** může být **lexikální**, **syntaktická**, **sémantická** a **referenční**
- lexikální – "stát," "žena," "hnát"
- syntaktická – "Jím špagety s masem."
  "Jím špagety se salátem."
  "Jím špagety s použitím vidličky."
  "Jím špagety se sebezapřením."
  "Jím špagety s přítelem."
- sémantická – "Jeřáb je vysoký." "Viděli jsme veliké oko."
- referenční – "Oni přišli pozdě." "Můžeš mi půjčit knihu?"
  "Ředitel vyhodil dělníka, protože (on) byl agresivní."

Anaforické a indexické výrazy

anaforické výrazy:

- *anaphora*
- používají **zájmena** pro odkazování na objekty zmíněné **dříve**
- "Poté co se Honza s Marií rozhodli se vzít, (oni) vyhledali kněze, aby je oddal."
- "Marie uviděla ve výloze prstýnek a požádala Honzu, aby jí ho koupil."

indexické výrazy:

- *indexicals*
- **odkazují** se na údaje v **jiných částech** promluvy
- "Já jsem tady."
- "Proč jsi to udělal?"

Metafora a metonymie

metafora:

- *metaphor*
- použití slov v **přeneseném významu** (na základě podobnosti), často systematicky
- "Zkoušel jsem ten proces zabít, ale nešlo to."
- "Bouře se vzteká."

metonymie:

- *metonymy*
- používání **jména** jedné **věci** pro (často zkrácené) označení **věci jiné**
- "Čtu Shakespeara."
- "Chrysler oznámil rekordní zisk."
- "Ten pstruh na másle u stolu 3 chce další pivo."

Nekompozicionalita

- *noncompositionality*
- příklady **porušení pravidla kompozicionality** u ustálených termínů nebo přednost jiného možného významu při určitých spojeních
- "aligátoří boty," "basketbalové boty," "dětské boty"
- "pata sloupu"
- "červená kniha," "červené pero"
- "bílý trpaslík"
- "dřevěný pes," "umělá tráva"
- "velká molekula"

Slovníky a encyklopedie

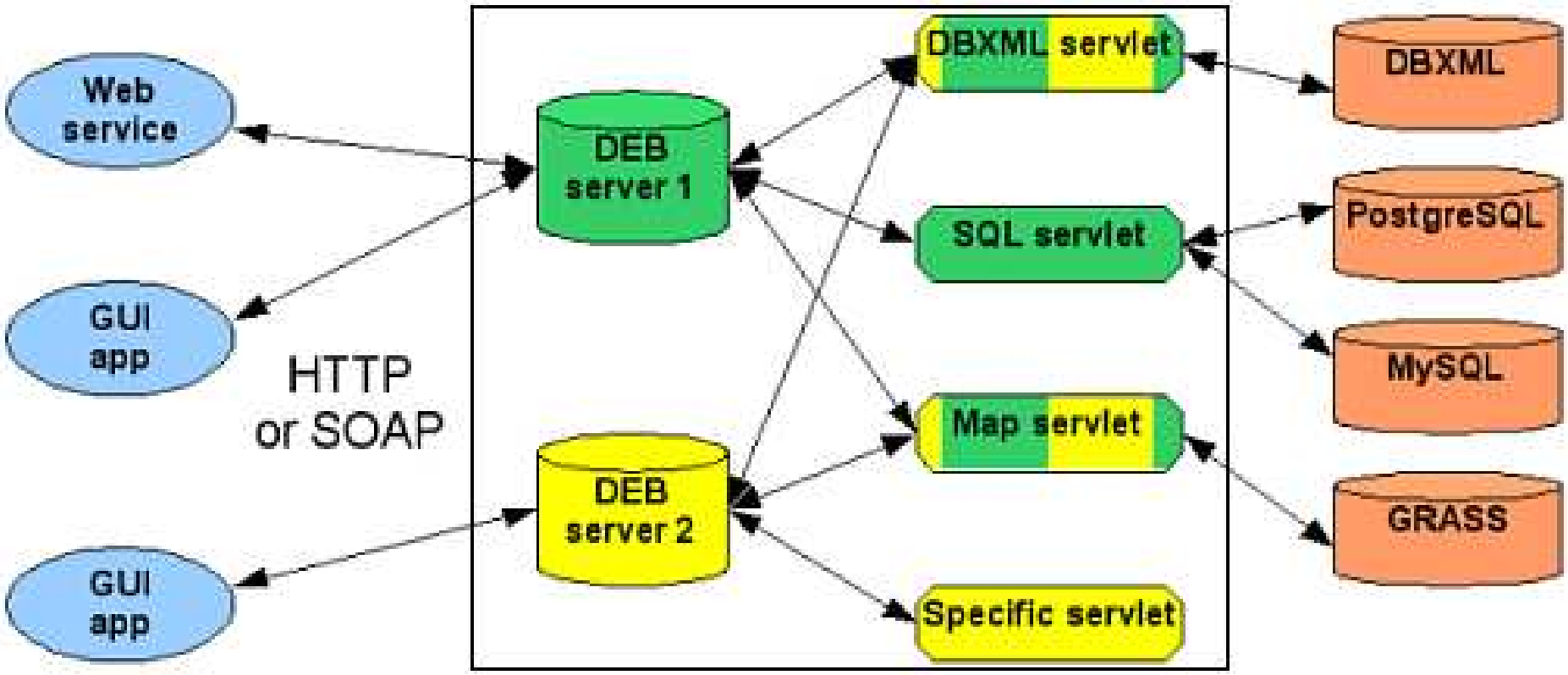
- specifikace **formy**:
  - grafická podoba – alternativy, dělení, velká počáteční písmena
  - zvuková podoba – výslovnost a její alternativy, slabiky, přízvuk, výška
- **gramatické** (morfo-syntaktické) **informace** – slovní druh a příslušné gramatické kategorie, morfologický vzor?
- specifikace **významu**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

slovník uvádí významy listémů, encyklopedie informace o jejich denotátech

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

DEB – platforma pro vývoj slovníků

- platforma pro vývoj systémů na psaní slovníků
  - `http://nlp.fi.muni.cz/projekty/deb2/`
  - 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*

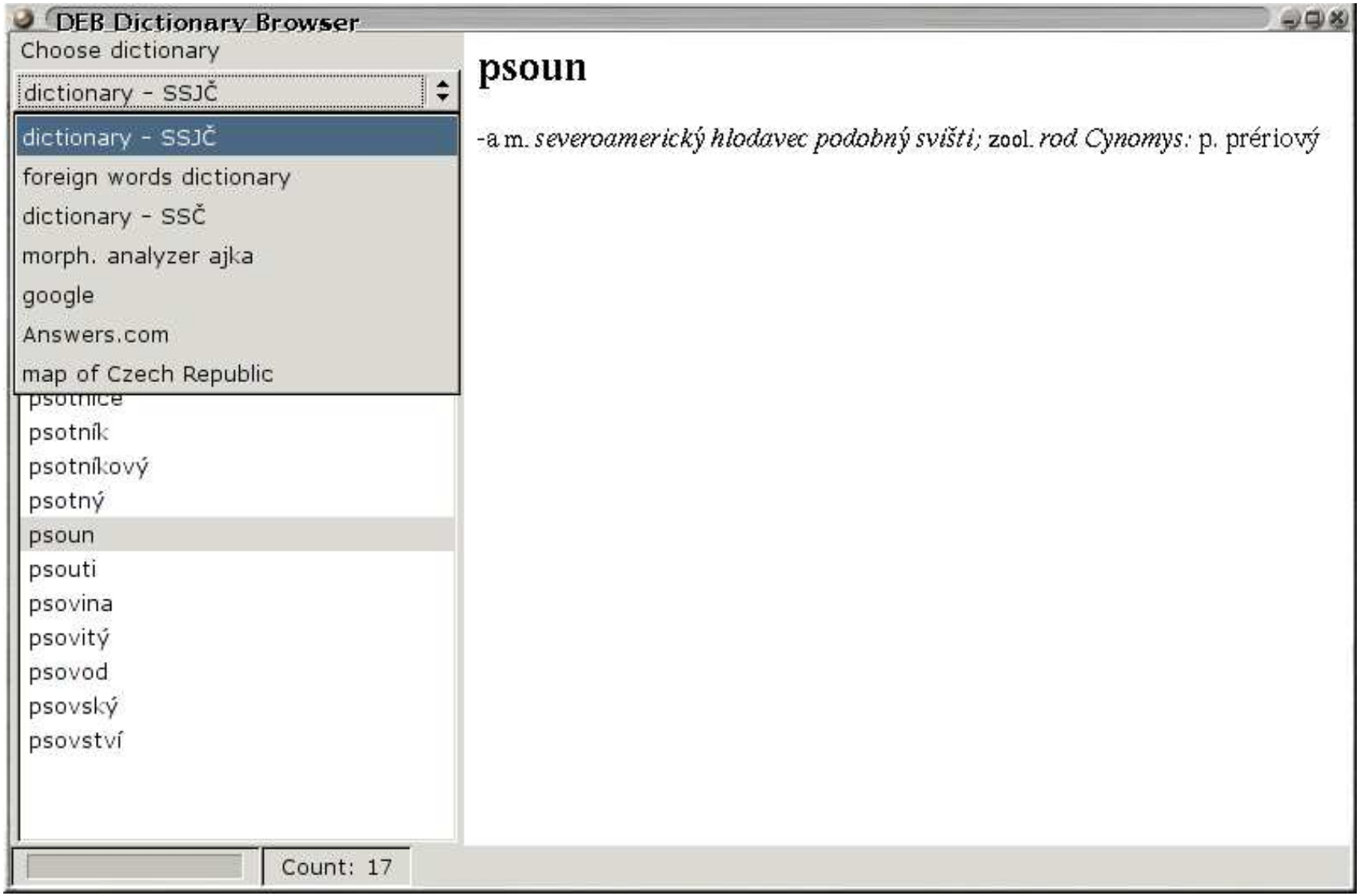


DEB používá komunikaci typu AJAX

DEBDict – příklad DEB klienta

jednoduchý klient původně určený pro demo základních funkcí

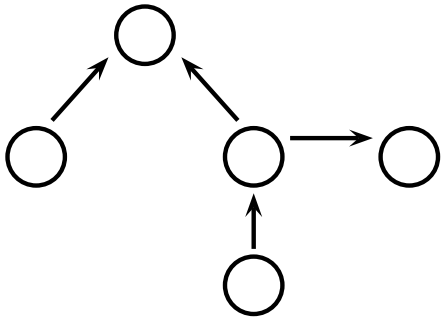
- dostupný jako instalovatelné **rozšíření Firefoxu** i jako vzdálená **webová služba**
- vícejazyčné uživatelské rozhraní (angličtina, čeština, další lze snadno doplnit)
- dotazy do několika XML slovníků s různou strukturou, výsledky jsou zpracovány XSLT transformací
- napojení na český morfologický analyzátor
- napojení na externí webové stránky (Google, Answers.com, Wikipedia)
- napojení na 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

Sémantické sítě

sémantické sítě – reprezentace faktových znalostí (pojmy + vztahy)

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



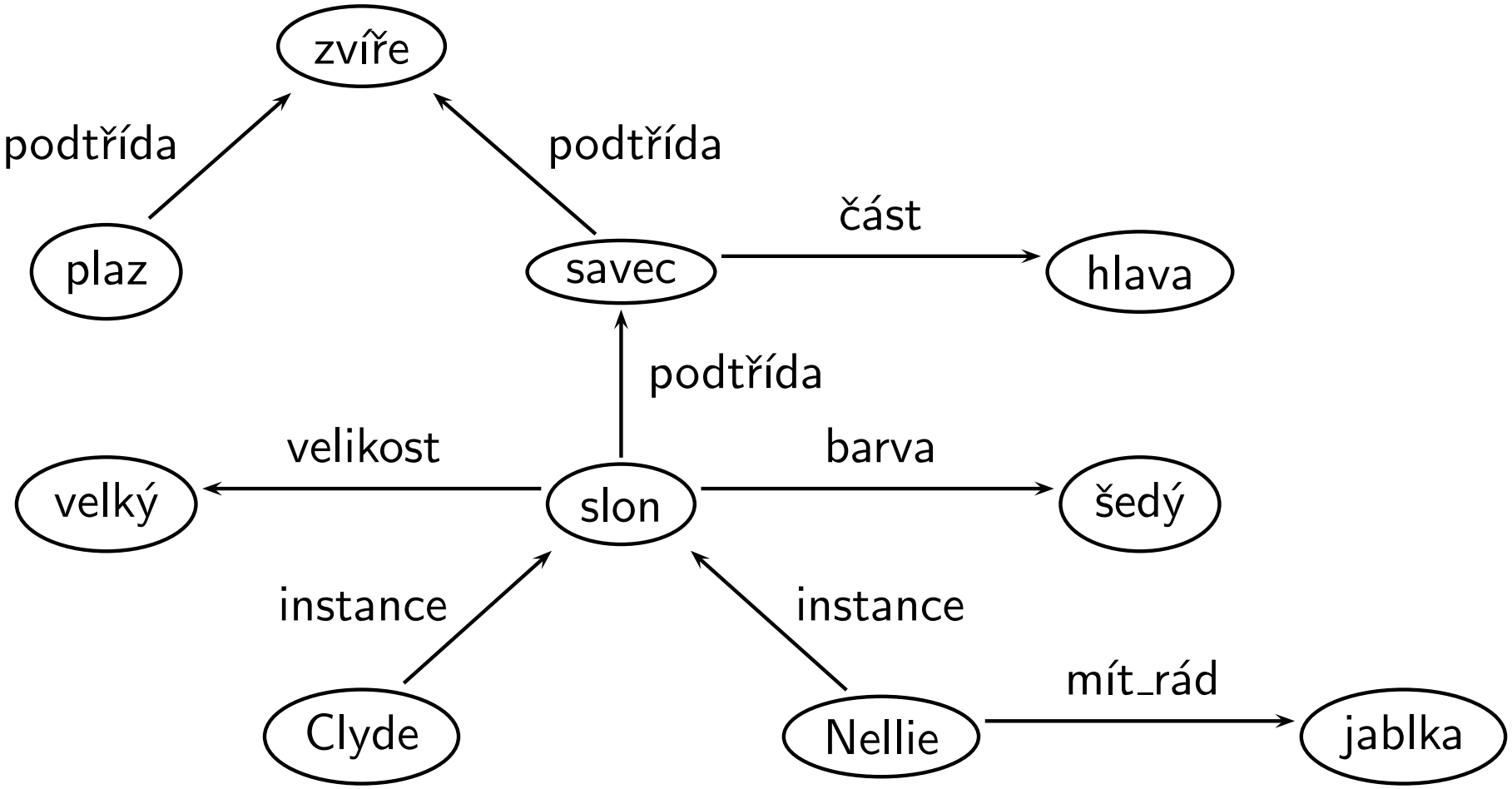
○ pojmy (objekty, třídy)

→ vztahy

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

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

Sémantické sítě – příklad



Dědičnost v sémantických sítích

- pojem sémantické sítě *předchazí* OOP
- **dědičnost**:
  - jestliže určitá vlastnost platí pro třídu→ platí i pro všechny její podtřídy
  - jestliže určitá vlastnost platí pro třídu→ platí i pro všechny prvky této třídy
- určení hodnoty vlastnosti – rekurzivní algoritmus
- potřeba specifikovat i výjimky – mechanizmus **vzorů** a **výjimek** (*defaults and exceptions*)
  - vzor – hodnota vlastnosti u třídy nebo podtřídy, platí ta, co je blíž objektu
  - výjimka – u konkrétního objektu, odlišná od vzoru

Dědičnost vztahů část/celek

- "krávy mají 4 nohy."
  - každá noha je částí krávy
- "Na poli je (konkrétní) kráva."
  - všechny části krávy jsou taky na poli
- "Ta kráva (na poli) je hnědá (celá)."
  - všechny části té krávy jsou hnědé
- "Ta kráva je šťastná."
  - ~~všechny části té krávy jsou šťastné~~ – neplatí
- lekce: některé vlastnosti jsou děděny částmi, některé nejsou
  explicitně se to vyjadřuje pomocí pravidel jako

$$\textit{part-of}(x, y) \wedge \textit{location}(y, z) \Rightarrow \textit{location}(x, z)$$

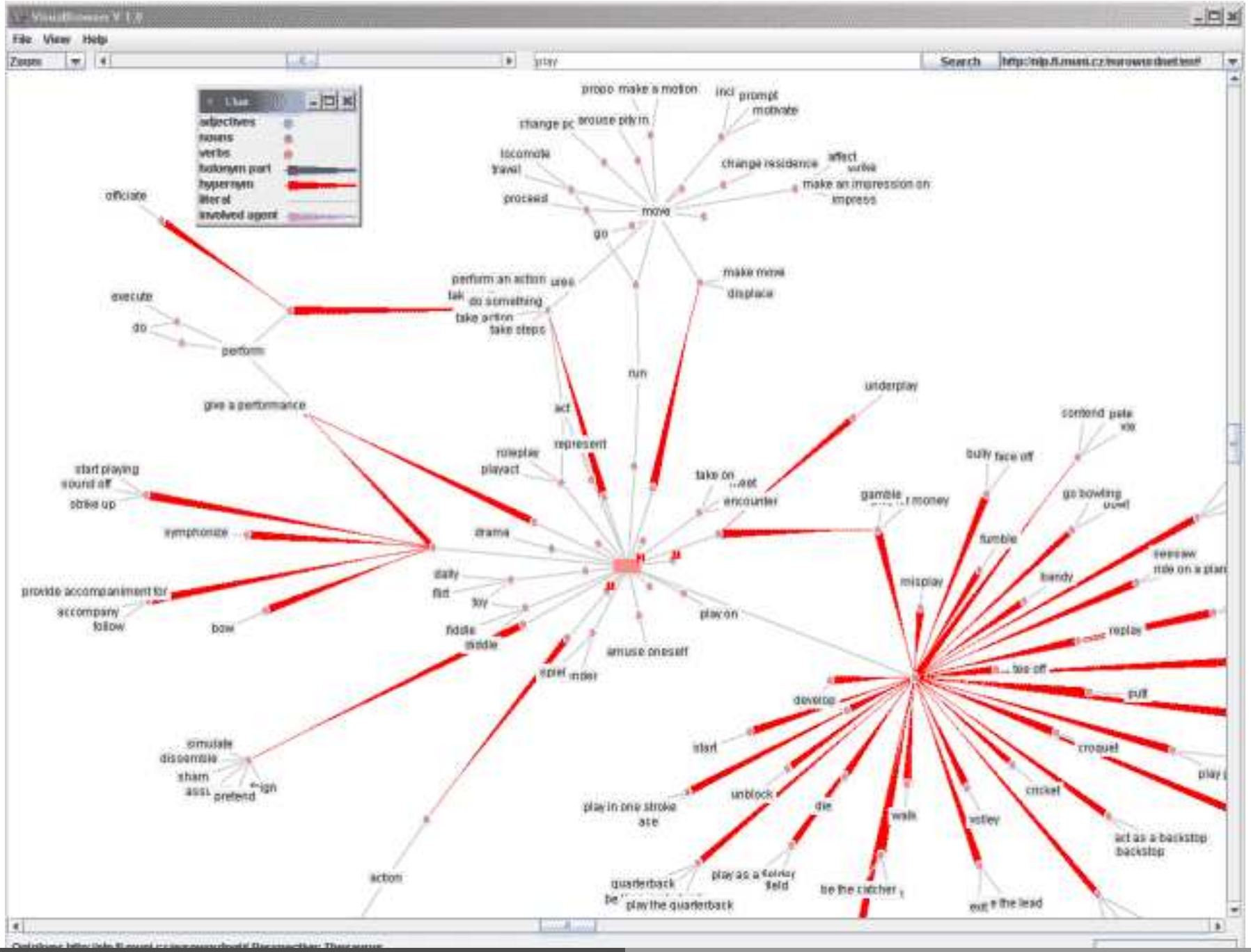
Vzory a výjimky – příklad

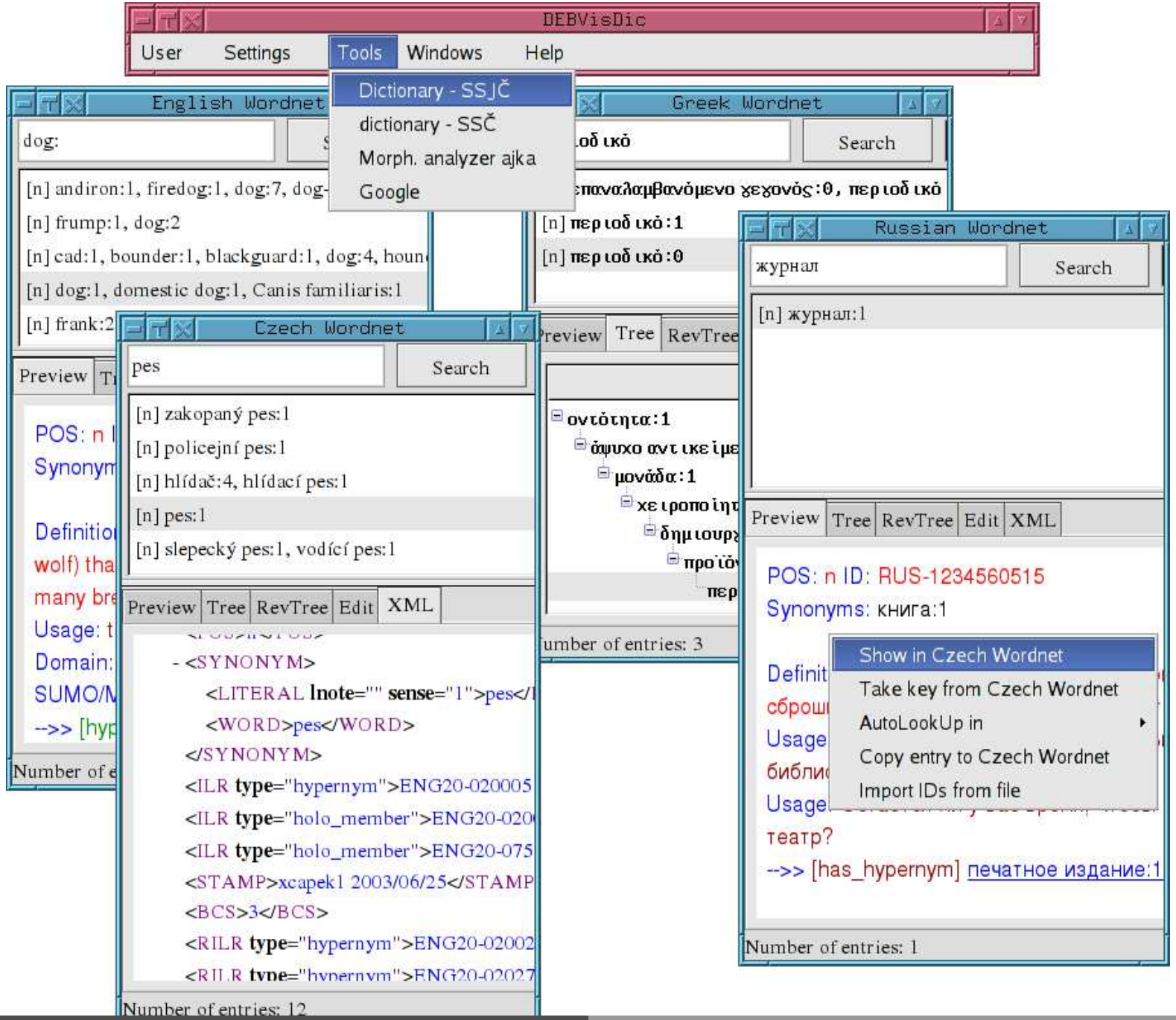
- "všichni ptáci mají křídla."
- "všichni ptáci umí létat."
- "ptáci se zlomenými křídly jsou ptáci, ale neumí létat."
- "tučnáci jsou ptáci, ale neumí létat."
- "kouzelní tučňáci jsou tučňáci, kteří umí létat."
- kdo umí létat:
  - "Tweety je pták."
  - "Petřík je tučnák."
  - "Penelope je kouzelný tučnák."
- všimněte si, že víra v hodnotu vlastnosti objektu se může měnit s příchodem nových informací o klasifikaci objektu

Aplikace sémantických sítí

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

- sématická síť 100.000 (anglických) pojmů, zachycuje:
  - synonyma, antonyma (významově stejná/opačná)
  - hyperonyma, hyponyma (podtřídy)
  - odvozenost a další jazykové vztahy
- tvoří se **národní wordnety** (navázané na anglický WN)
  český wordnet – cca 30.000 pojmů
- nástroj na editaci národních wordnetů – VisDic, 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
play
objectives
nouns
verbs
holonym part
hypernym
literal
involved agent
officiate
execute
do
perform
give a performance
start playing
sound off
strike up
symphonize
provide accompaniment for
accompany
follow
bow
simulate
dissemble
sham
assi
pretend
action
perform an action
do something
take action
take steps
move
go
make move
displace
run
act
represent
roleplay
playact
drama
daily
flirt
toy
fiddle
middle
amuse oneself
play on
take on
encounter
underplay
gamble
money
bully
face off
fumble
misplay
handy
replay
pull
croquet
cricket
act as a backstop
backstop
walk
volley
be the catcher
play the quarterback
quarterback
play as a
field
die
unblock
start
develop
play in one stroke
ate
make an impression on
impress
change residence
affect
change pc
travel
process
make a motion
prompt
motivate


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
οντότητα:1
μονάδα:1
Number of entries: 3
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
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

České valenční lexikony

zdroje (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 aktuálně se do něj doplňují slovesa z Briefu
- edituje se 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 | wn link | verb class | functors | forms | aspect | complexity | miscel. | search | home | help

- A (18)
- B (101)
- C (11)
- Č (18)
- D (457)
- E (6)
- F (11)
- H (68)
- CH (34)
- I (8)
- J (14)
- K (70)
- L (24)
- M (64)
- N (249)
- O (315)
- P (572)
- R (84)
- Ř (42)
- S (217)
- Š (33)
- T (25)
- U (160)
- V (469)
- Z (368)
- Ž (29)

- tahat$_1$
- tahat$_2$
- táhnout$_3$
- táhnout$_6$
- táhnout se$_1$
- téci$_1$
- téci$_1$
- téct$_1$
- téct$_1$
- teoretizovat$_1$
- testovat$_1$
- těžit$_2$
- těžit$_3$
- tisknout$_2$
- tlačit$_2$
- tlačit$_2$
- tlačit$_3$
- tlouct se$_1$
- toulat se

dobývat$^1_{impf}$ / těžit$^2_{impf}$ / lámat$^3_{impf}$

1 dobývat$_1$ / těžit$_2$ / lámat$_3$ ≈

-frame: AG<person:1> obl kdo1 VERB obl SUBS<substance:1> obl co4

-example: **ned:** *lámal v dolech kámen*

-synonym:

-use: prim

2 dobývat$_1$ / těžit$_2$ / lámat$_3$ ≈

-frame: AG<institution:1> obl co1 VERB obl SUBS<substance:1> obl co4

-example: **ned:** *tato společnost těží mramor*

-synonym:

-use: prim

Samostatný úkol – VerbaLex

- veškerá práce probíhá na serveru laboratoře NLP – `aurora.fi.muni.cz`
- krátký fragment Briefu ve formátu VerbaLexu
  editace příkazem '`/nlp/verbalex/bin/verbalex.sh IB030`'
- doplnit všechny údaje valenčních rámců pro uvedená slovesa
- návody viz `http://nlp.fi.muni.cz/verbalex/` → HTML → Help + v něm uvedené odkazy
- s využitím zdrojů:
  - VerbaLex – `http://nlp.fi.muni.cz/verbalex/`
  - DEBDict – spuštění ve Firefoxu, viz předchozí popis DEB platformy
  - VisDic – příkaz '`visdic`'
  - korpus SYN2000 (Bonito, Word Sketches) – `http://corpora.fi.muni.cz/syn2k/`, login/heslo = `ib030`/`ib030`