

# Sémantika a základní sémantické reprezentace

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:

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

**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

**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

**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, slovtvorba
- logická sémantika – výroková, predikátová a vyšší logiky
- lingvistická pragmatika

*entail* = znamenat, vyplývat; nutnost a očekávanost

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

# Význam v jazyce

Rozdělení studia významu v jazyce:

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

*entail* = znamenat, vyplývat; nutnost a očekávanost

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


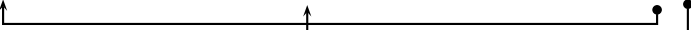


# Víceznačnost

- *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:


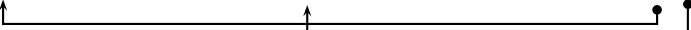
- *anaphora*
- používají **zájmena** pro odkazování na objekty zmíněné **dříve**
- “Poté co se Honza s Marií rozhodli se vzít, (**oni**) vyhledali kněze, aby **je** oddal.”  

- “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?”

# 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.”

# 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”

# Obsah

- 1 Sémantika
  - Význam v jazyce
  - Problémy při analýze přirozeného jazyka
- 2 Slovníky a encyklopedie
  - DEB – platforma pro vývoj slovníků
- 3 Sémantické sítě
  - Sémantické sítě
- 4 Reprezentace slovesných valencí
  - České valenční lexikony
  - Valenční lexikon VerbaLex

# 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)



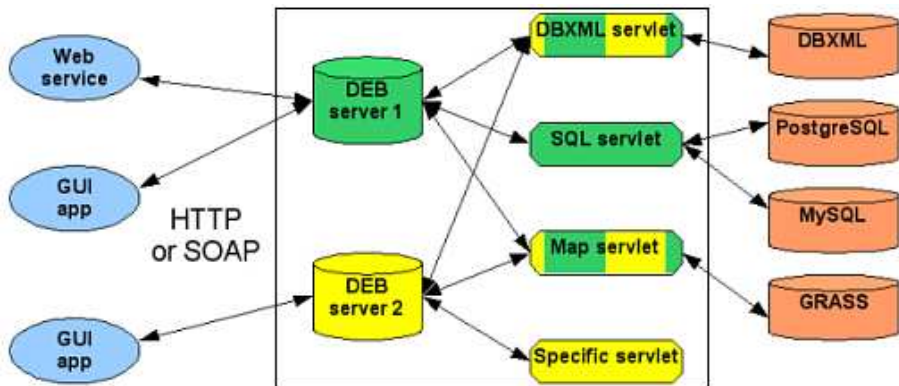
# Slovníky a encyklopedie

- 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*

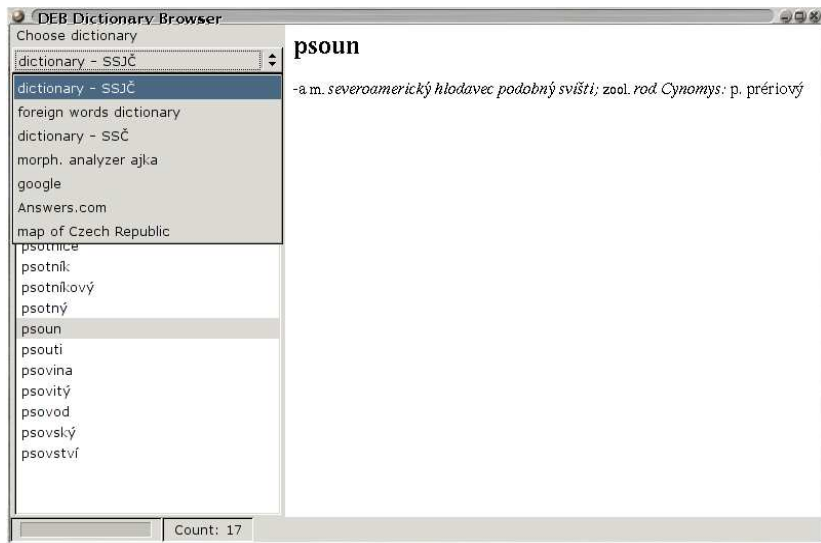


DEB používá komunikaci typu AJAX

# 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



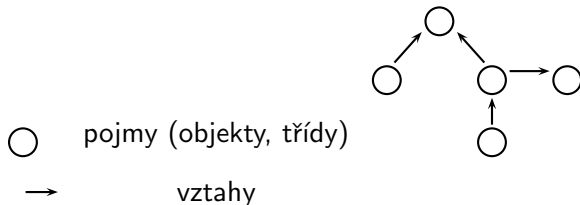
# Obsah

- 1 Sémantika
  - Význam v jazyce
  - Problémy při analýze přirozeného jazyka
- 2 Slovníky a encyklopedie
  - DEB – platforma pro vývoj slovníků
- 3 Sémantické sítě
  - Sémantické sítě
- 4 Re prezentace slovesných valencí
  - České valenční lexikony
  - Valenční lexikon VerbaLex

# Sémantické sítě

**sémantické sítě** – reprezentace faktových znalostí (pojmy + vztahy)

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

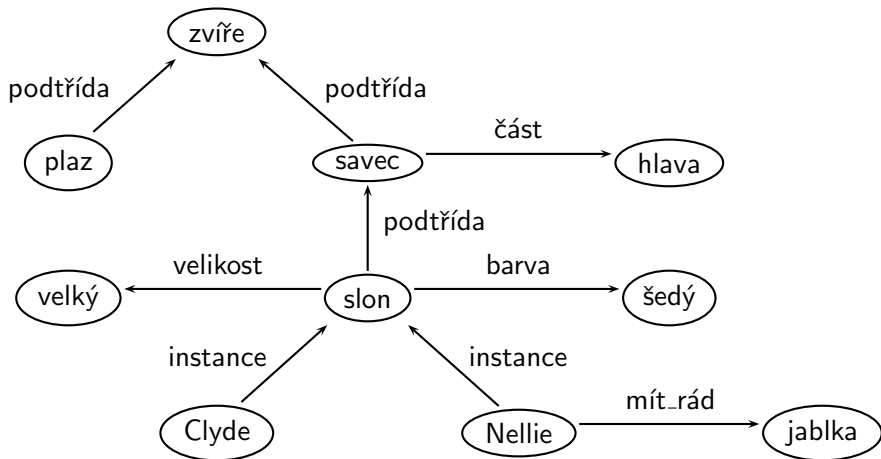


- nejdůležitější vztahy:

- **podtřída** (*subclass*) – 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 *předchází* 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 – mechanismus **vzorů** a **výjimek** (*defaults and exceptions*)
  - vzor – hodnota vlastnosti u třídy nebo podtřídy, platí ta, co je blíže 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

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

# 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

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

# Vzory a výjimky – příklad

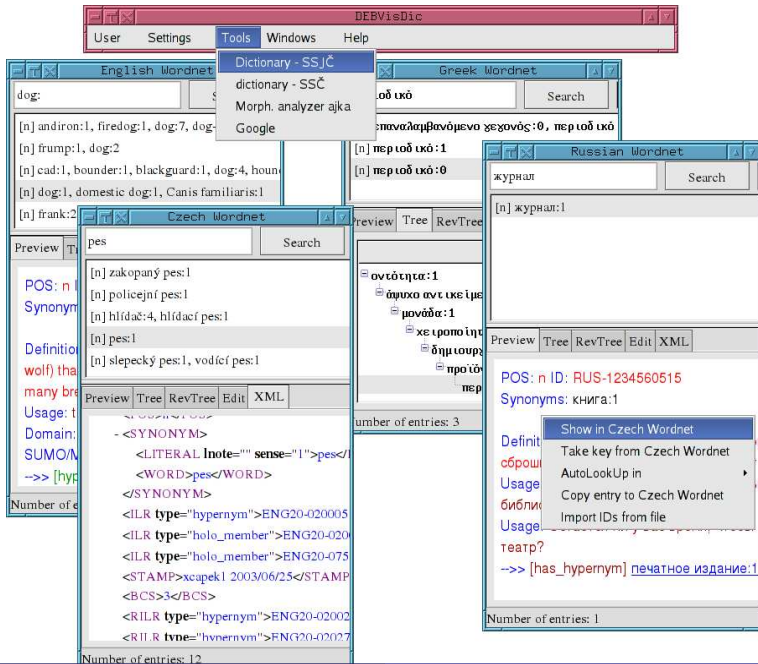
- “všichni ptáci mají křídla.”
- “všichni ptáci umí létat.”
- “ptáci se zlomenými křídly jsou ptáci, ale neumí létat.”
- “tučnáci jsou ptáci, ale neumí létat.”
- “kouzelní tučňáci jsou tučňáci, kteří umí létat.”
- kdo umí létat:
  - “Tweety je pták.”
  - “Petřík je tučnák.”
  - “Penelope je kouzelný 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/) – <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 – DEBVisDic, vyvinutý na FI MU
- VisualBrowser –  
<http://nlp.fi.muni.cz/projekty/visualbrowser/>  
nástroj na vizualizaci (sémantických) sítí, vznikl jako DP na FI MU





# Obsah

- 1 Sémantika
  - Význam v jazyce
  - Problémy při analýze přirozeného jazyka
- 2 Slovníky a encyklopedie
  - DEB – platforma pro vývoj slovníků
- 3 Sémantické sítě
  - Sémantické sítě
- 4 Reprezentace slovesných valencí
  - České valenční lexikony
  - Valenční lexikon VerbaLex



# 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)
```

# 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, 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: lámat
```

```
+ ACT(1;obl) PAT(4;obl)
```

# 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
- experty 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<sub>1</sub>
- tahat<sub>2</sub>
- táhnout<sub>3</sub>
- táhnout<sub>6</sub>
- táhnout se<sub>1</sub>
- téci<sub>1</sub>
- téci<sub>1</sub>
- téct<sub>1</sub>
- téct<sub>1</sub>
- teoretizovat<sub>1</sub>
- testovat<sub>1</sub>
- těžít<sub>2</sub>
- těžít<sub>3</sub>
- tisknout<sub>2</sub>
- tlačit<sub>2</sub>
- tlačit<sub>2</sub>
- tlačit<sub>3</sub>
- tlouct se<sub>1</sub>
- toulat se

**dobývat**<sup>1</sup> / **těžít**<sup>2</sup> / **lámat**<sup>3</sup>

1 dobývat<sub>1</sub> / těžít<sub>2</sub> / lámat<sub>3</sub> =

-frame: **AG**<person:1><sub>kdo1</sub> **VERB**<sup>obl</sup> **SUBS**<substance:1><sub>obl co4</sub>

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

-synonym:

-use: prim

2 dobývat<sub>1</sub> / těžít<sub>2</sub> / lámat<sub>3</sub> =

-frame: **AG**<institution:1><sub>co1</sub> **VERB**<sup>obl</sup> **SUBS**<substance:1><sub>obl co4</sub>

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

-synonym:

-use: prim

# Samostatný úkol – VerbaLex

- veškerá práce probíhá na serveru laboratoře NLP – `aurora.fi.muni.cz`, kromě DEB aplikací (jdou i vzdáleně)
- 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
  - DEBVisDic – instalace a následné spuštění ve Firefoxu  
instalační balíky pro DEBDict i DEBVisDic jsou na `http://nlp.fi.muni.cz/projects/deb2/clients/`  
přístup pro DEBDict i DEBVisDic je přes `login/heslo = ib030/ib030`
  - korpus SYN2000 (Bonito, Word Sketches) – `http://corpora.fi.muni.cz/syn2k/`, `login/heslo = ib030/ib030`

# Samostatný úkol – NLTK

jiná možnost úkolu - nutné nahlásit předem:

- **NLTK**, **Natural Language Toolkit**, [nltk.sourceforge.net](http://nltk.sourceforge.net)
- naprogramování ukázkového příkladu algoritmu z přednášky
- algoritmus musí být předveden na datech pro češtinu
- je možné i převzít hotový ukázkový příklad a upravit jej pro češtinu
- stačí krátký, ale smysluplný "výukový" příklad
- nutné nahlásit **do 8.5. e-mailem** na [hales@fi.muni.cz](mailto:hales@fi.muni.cz) včetně stručného popisu zvoleného algoritmu, tato volba musí být schválena přednášejícím
- při nenahlášení NLTK úkolu se předpokládá řešení VerbaLex úkolu
- obě varianty úkolu musí být odevzdány (tj. poslána informace o dokončení úkolu e-mailem na [hales@fi.muni.cz](mailto:hales@fi.muni.cz) se **Subject: IB030 – odevzdani ukolu**) **před termínem zkoušky**