

Reprezentace znalostí a základní sémantické struktury

Aleš Horák

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

Obsah:

- ▶ Reprezentace znalostí
- ▶ Sémantické datové struktury
- ▶ Slovníky a specializované lexikony

otázka:

Jak zapíšeme znalosti o problému/doméně?

Když je zapíšeme, můžeme z nich mechanicky odvodit nová fakta?

- ▶ **reprezentace znalostí** (*knowledge representation*) – hledá způsob vyjádření znalostí počítačově zpracovatelnou formou (za účelem odvozování)
- ▶ **vyvozování znalostí** (*reasoning*) – zpracovává znalosti uložené v **bázi znalostí** (*knowledge base, KB*) a provádí **odvození** (inference) nových závěrů:
 - odpovědi na dotazy
 - zjištění faktů, které vyplývají z faktů a pravidel v KB
 - odvodit akci, která vyplývá z dodaných znalostí, ...

Reprezentace znalostí

proč je potřeba speciální **reprezentace znalostí**?

vnímání lidí × *vnímání počítačů*

▶ člověk

- ▶ když dostane novou věc (třeba pomeranč) – **prozkoumá** a **zapamatuje** si ho (a třeba sni)
- ▶ během tohoto procesu člověk zjistí a uloží všechny základní vlastnosti
- ▶ později, když se **zmíní** daná věc, vyhledají se a připomenou uložené informace

▶ počítač

- ▶ musí se spolehnout na informace od lidí
- ▶ jednodušší informace – přímé *programování*
- ▶ složité informace – zadané v **symbolickém jazyce**

Volba reprezentace znalostí

která **reprezentace znalostí** je **nejlepší**?

Pro řešení skutečně obtížných problémů musíme použít několik různých reprezentací. Důvodem pro to je to, že každý typ datových struktur má své přínosy i nedostatky a žádná z nich není adekvátní pro všechny různé funkce používané v tom, čemu říkáme “zdravý rozum” (common sense).

– Marvin Minsky

Reprezentace znalostí pomocí logiky nebo datových struktur

Logika:

- ▶ znalosti uloženy ve formě **logických formulí**
- ▶ vyvozování nových znalostí = hledání **důkazu**

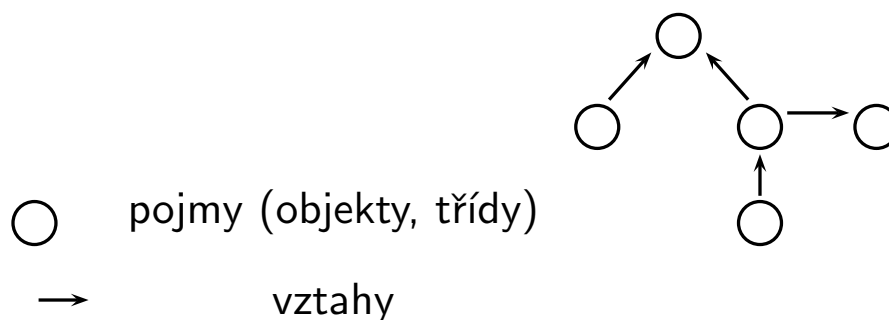
Specializované datové struktury:

- ▶ sémantické sítě
- ▶ rámce
- ▶ pravidlové systémy
- ▶ struktury pro práci s nejistotou a pravděpodobností

Sémantické sítě

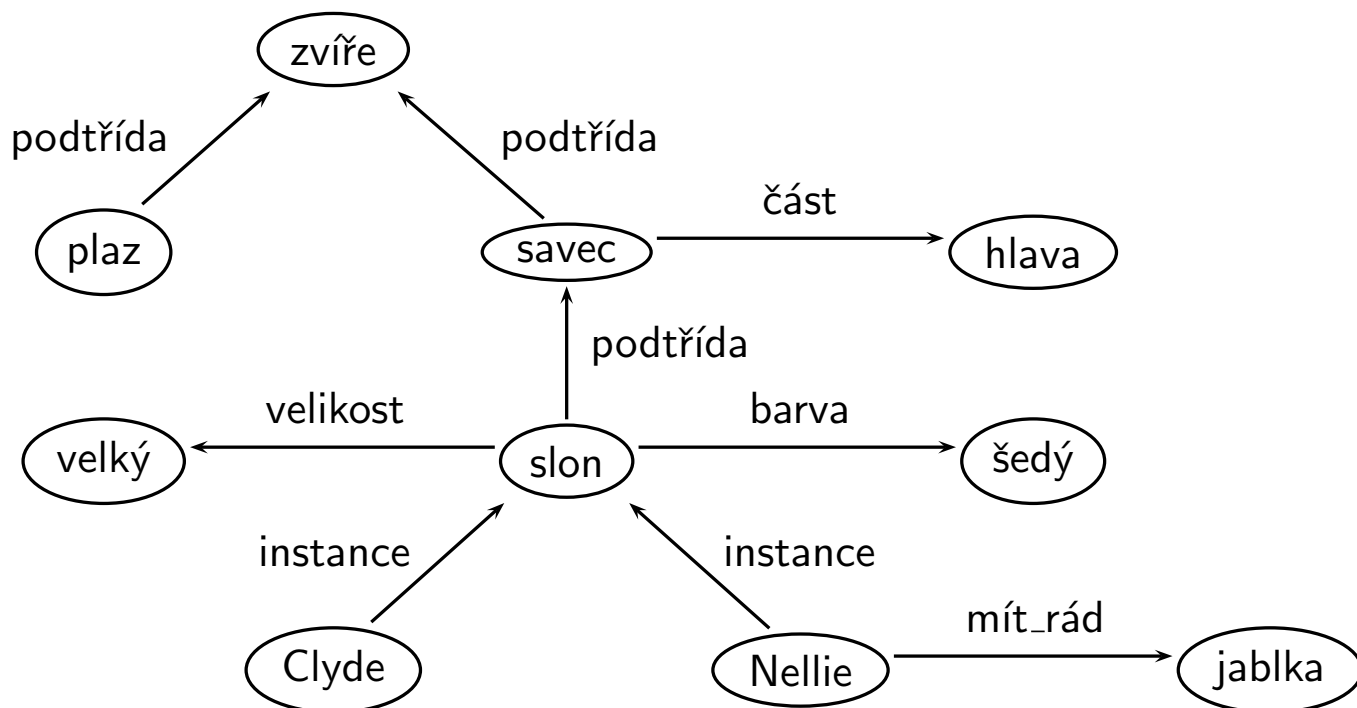
sémantické sítě – reprezentace faktových znalostí (pojmy + vztahy)

- ▶ 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

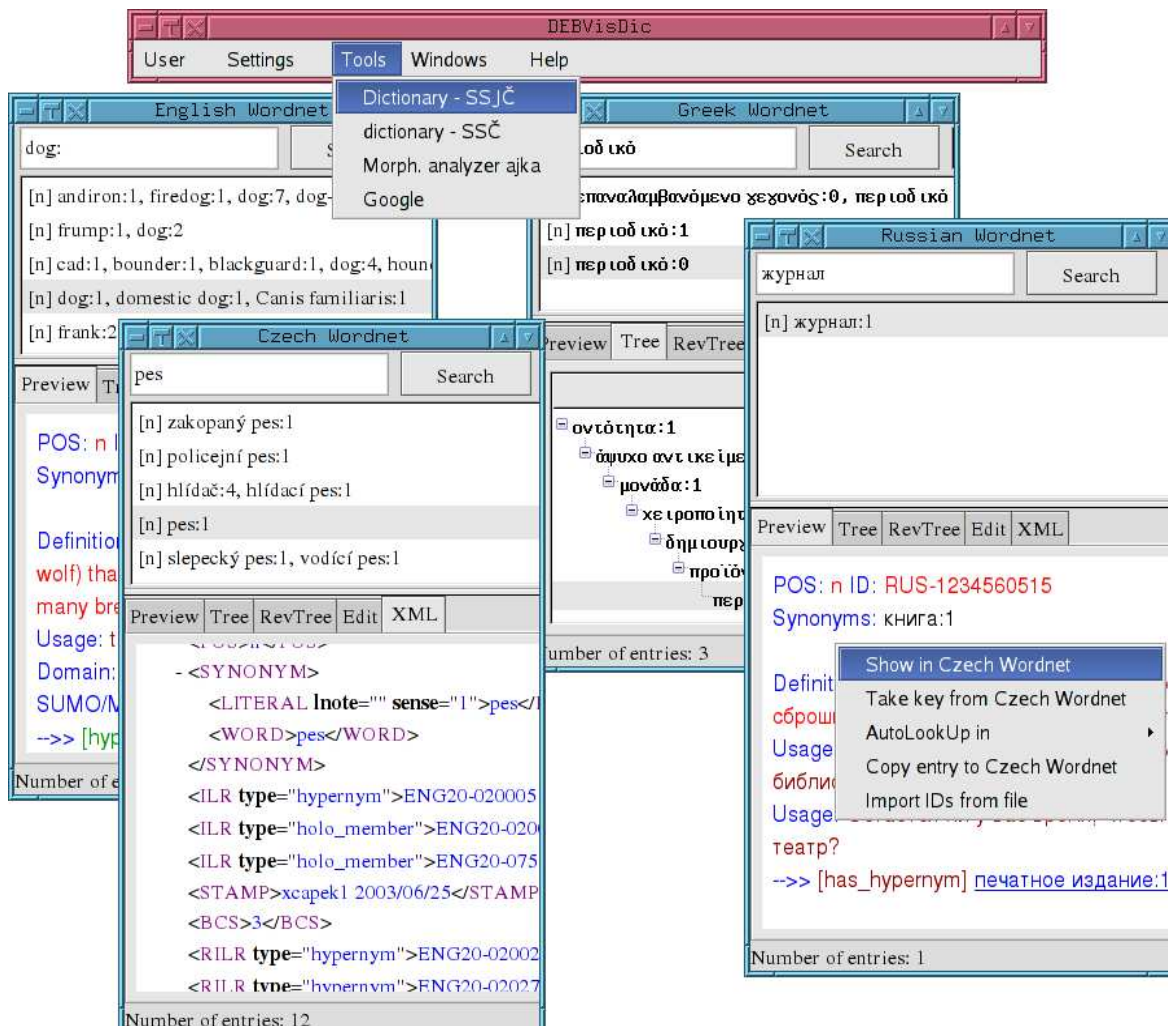
Dědičnost vztahů část/celek

- ▶ “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

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

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:
 - “Penelope je pták.” \Rightarrow “Penelope **umí** létat”
 - “Penelope je tučnák.” \Rightarrow “Penelope **neumí** létat”
 - “Penelope je kouzelný tučnák.” \Rightarrow “Penelope **umí** létat”
- ▶ všimněte si, že víra v hodnotu vlastnosti objektu se může měnit s příchodem nových informací o klasifikaci objektu



Rámce

Rámce (frames):

- ▶ varianta sémantických sítí
- ▶ velice populární pro reprezentaci znalostí v expertních systémech
- ▶ všechny informace relevantní pro daný pojem se ukládají do univerzálních struktur – **rámců**
- ▶ stejně jako sémantické sítě, rámce podporují dědičnost
- ▶ OO programovací jazyky vycházejí z teorie rámců

Rámce – příklad

rámec obsahuje **objekty**, *sloty* a hodnoty slotů
příklady rámců:

savec:

podtřída: zvíře
část: hlava
**má_kožich:* ano

slon:

podtřída: savec
**barva:* šedá
**velikost:* velký

Nellie:

instance: slon
mít_rád: jablka

'*' označuje **vzorové hodnoty**, které mohou měnit hodnoty u podtříd a instancí

Sémantické sítě × rámce

sémantické sítě	rámce
uzly	objekty
spoje	sloty
uzel na druhém konci spoje	hodnota slotu

deskripční logika – logický systém, který manipuluje přímo s rámci

Pravidlové systémy

- ▶ snaha zachytit **produkčními pravidly** znalosti, které má expert
- ▶ obecná forma pravidel

IF *podmínka*
THEN *akce*

- podmínky – booleovské výrazy, dotazy na hodnoty **proměnných**
- akce – nastavení hodnot proměnných, příznaků, ...
- ▶ důležité vlastnosti:
 - znalosti mohou být strukturovány do modulů
 - systém může být snadno rozšířen přidáním nových pravidel beze změny zbytku systému

Metody pro práci s nejistotou

definujme akci A_t jako “Vyrazit na letiště t hodin před odletem letadla.”
jak najít odpověď na otázku “*Dostanu se akcí A_t na letiště včas?*”

▶ defaultní/nemonotónní logika

Předpokládejme, že nepíchnu cestou kolo.

Předpokládejme, že A_5 bude OK, pokud se nenajde protipříklad.

▶ pravidla s faktory nejistoty

$A_5 \mapsto_{0.3}$ dostat se na letiště včas.

zalévání $\mapsto_{0.99}$ mokrý trávník

mokrý trávník $\mapsto_{0.7}$ déšť

▶ pravděpodobnost

Vzhledem k dostupným informacím, A_3 mě tam dostane včas s pravděpodobností 0.05.

Použití **náhodných proměnných** a pravidel pro výpočet pravděpodobnosti logicky souvisejících událostí (podmíněná pravděpodobnost, bayesovské pravidlo, ...)

Slovníky a specializované lexikony

Slovníky typicky obsahují:

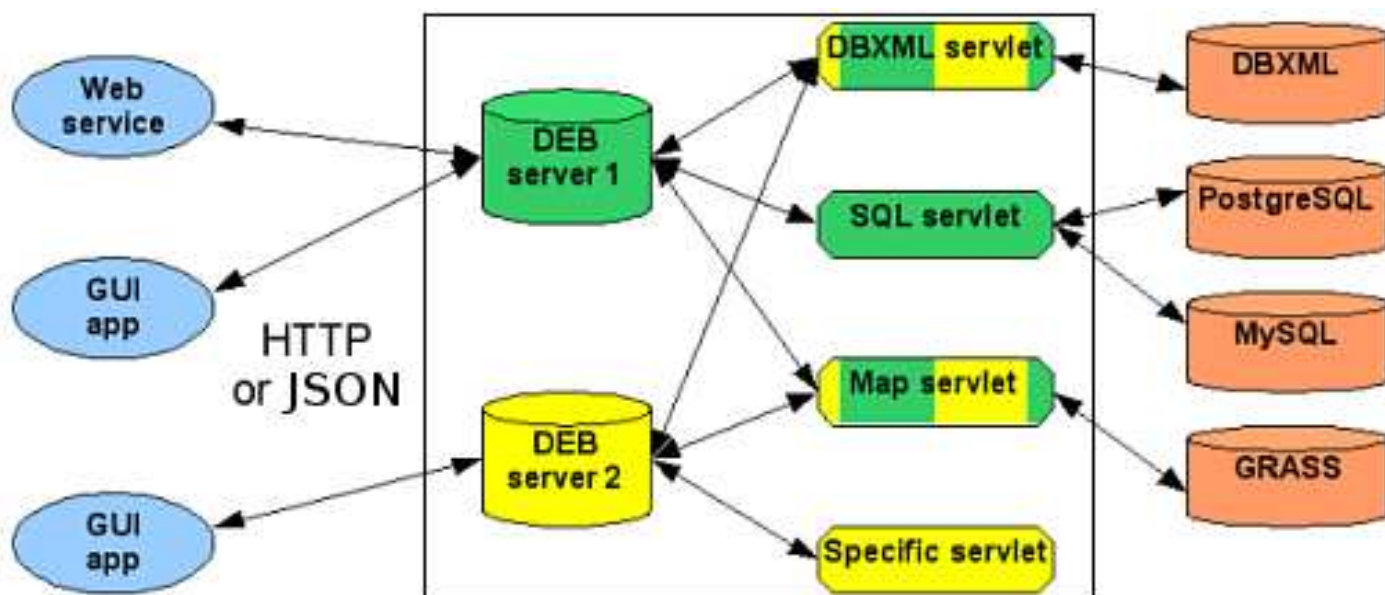
- ▶ specifikace **formy**:
 - grafická podoba – alternativy, dělení, velká počáteční písmena
 - zvuková podoba – výslovnost a její alternativy, slabiky, přízvuk, výška
- ▶ **gramatické** (morfo-syntaktické) **informace** – slovní druh a příslušné gramatické kategorie, morfologický vzor?
- ▶ specifikace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

slovník uvádí významy listémů, **encyklopedie** informace o jejich denotátech

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

DEB – platforma pro vývoj slovníků

- ▶ platforma pro vývoj systémů na psaní slovníků
 - <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*

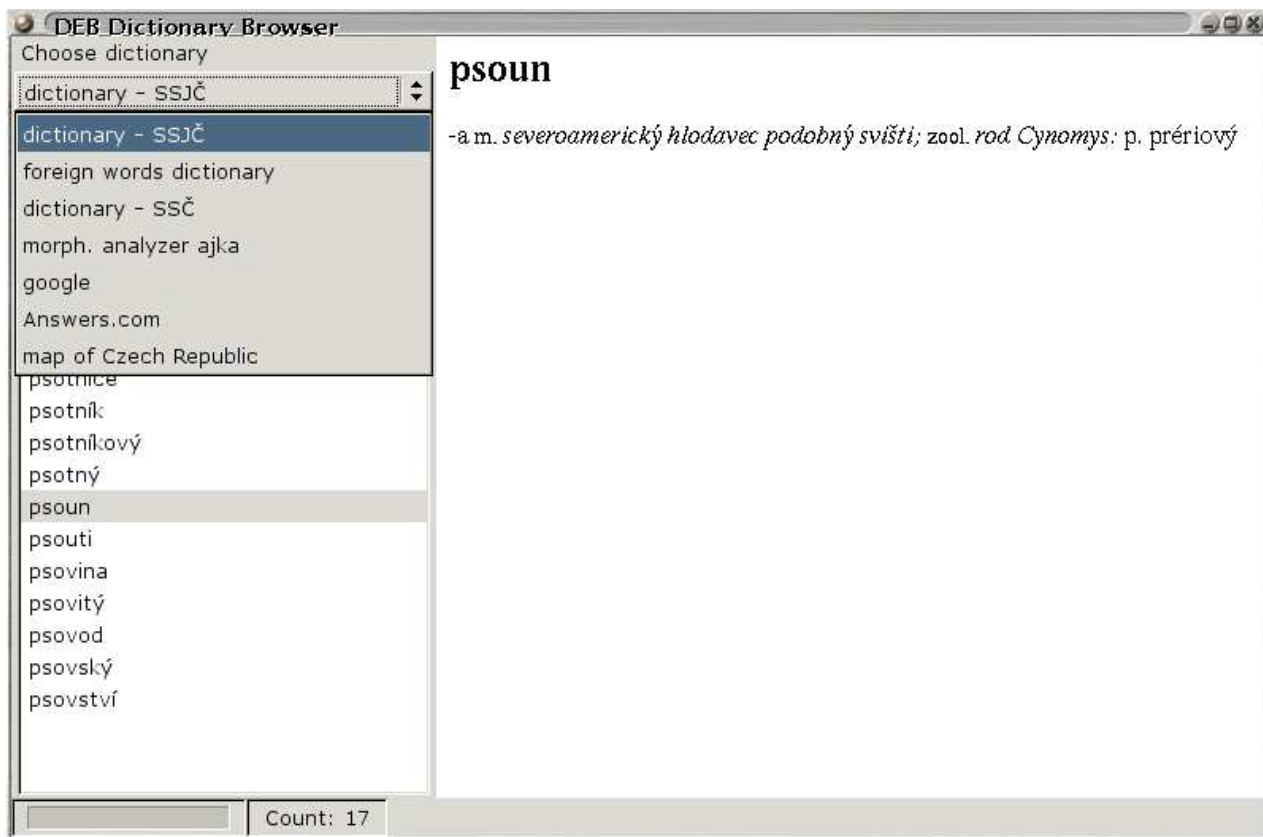


DEB používá komunikaci typu AJAX

DEBDict – příklad DEB klienta

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



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, 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)

Valeční lexikon VerbaLex

- ▶ vznikl na začátku roku 2005, využívá všech dostupných zdrojů
- ▶ 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
- ▶ exporthy z XML do HTML pro prohlížení a PDF pro tisk

VerbaLex v HTML

The screenshot shows the VerbaLex web interface. At the top, there are navigation tabs: 'alphabet', 'wn link', 'verb class', 'functors', 'forms', 'aspect', 'complexity', 'miscel.', a search box, and buttons for 'home' and 'help (?)'. Below the tabs is a list of letters from A to Ž with the number of entries for each. The main content area is divided into three columns. The left column shows a list of verbs with their grammatical forms (e.g., 'tahat₁', 'tahat₂', 'táhnout₃', etc.). The middle column shows the selected verb 'dobývat¹ / těžit² / lámat³' with its grammatical forms and semantic roles. The right column shows the detailed information for the selected verb, including its frame, example, synonyms, and use.

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)

tahat₁
tahat₂
táhnout₃
táhnout₆
táhnout se₁
téci₁
téci₁
téct₁
téct₁
teoretizovat₁
testovat₁
těžit₂
těžit₃
tisknout₂
tlačit₂
tlačit₂
tlačit₃
tlouct se₁
toulat se

dobývat¹ / těžit² / lámat³
 1 dobývat₁ / těžit₂ / lámat₃ =
 -frame: **AG**<person:1>_{kdo1} **VERB**^{obl} **SUBS**<substance:1>_{obl co4}
 -example: ned: lámal v dolech kámen
 -synonym:
 -use: prim

2 dobývat₁ / těžit₂ / lámat₃ =
 -frame: **AG**<institution:1>_{obl co1} **VERB**^{obl} **SUBS**<substance:1>_{obl co4}
 -example: ned: tato společnost těží mramor
 -synonym:
 -use: prim

Využití valencí v sémantické analýze

reprezentace **slovesného rámce**:

1. syntaktické rysy:

dávat něco_{neživ.NP}, 4.pád, bez předložky

někomu_{živ.NP}, 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 / $(o(o\pi)(o\pi))_{\omega ll} \quad x \dots l \quad y \dots l : s_{wt}y, s \dots (ol)_{\tau\omega}$

překlad z valenčního výrazu do funkce významu:

typ argumentu = typ	}	<ul style="list-style-type: none"> ▶ jmenné skupiny ▶ příslovečné fráze ▶ vedlejší věty ▶ infinitivu
---------------------	---	------------------------------------------------------------------------------------------------------------------------------------------------