Pala, Karel; Ševeček, Pavel

Česká lexikální databáze typu WordNet

Sborník prací Filozofické fakulty brněnské univerzity. A, Řada jazykovědná. 1999, vol. 48, iss. A47, pp. [51]-64

ISBN 80-210-2098-9
ISSN 0231-7567

SBORNÍK PRACÍ FILOZOFICKÉ FAKULTY BRNĚNSKÉ UNIVERZITY
STUDIA MINORA FACULTATIS PHILOSOPHICAE UNIVERSITATIS BRUNENSIS
A 47, 1999

KAREL PALA, PAVEL ŠEVEČEK

ČESKÁ LEXIKÁLNÍ DATABÁZE TYPU WORDNET
(V RÁMCI PROJEKTU EUROWORDNET-2)

1. Úvod – motivace

Standardní způsobem organizace lexikálního materiálu ve slovnících je abecední řazení (lexikografické uspořádání). Hledání v abecedně řazených slovnících hledání je pomalé, i když počítače nyní umožňují prohlížení zrychlit. Je však zjevně neefektivní užívat počítačů jen jako „obracečů" stránek a má smysl hledat vhodnější způsoby organizace slovníku. Položme si otázku, zda v tomto ohledu existuje cesta vedoucí ke zlepšení dosavadních standardních slovníků.

Příklady ukazují, že třeba u lexikální jednotky *strom* s významem rostlina najdeme následující definici: *dřevina s kmenem, který se nahoře větví v korunu: listnaté, jehličnaté, ovocné...* (SSČ, 1994, s. 419). Jako u většiny definic ve standardních slovnících je i zde použito základní schéma: **genus proximum plus rozlišující příznaky popisující specifické rysy stromu** (a obvykle mající formu vztažné věty). Z pohledu běžného uživatele v definici nic nechybí, ale nicméně nezmiňuje se o tom, že stromy mají kořeny, skládají se z buněk nebo že jsou to živé organismy. Informaci tohoto druhu ale můžeme najít u nadřazeného výrazu *rostlina.* Dále, definice výrazu *strom* neobsahuje informaci o jiných podobných typech rostlin, tedy o třeba o keřích. Každý uživatel slovníku dobře ví, že najít ve standardním slovníku informace o lexikálních jednotkách stejného typu je časově velmi náročné. V podobné situaci je uživatel standardního slovníku, když se chce něco dovědět o jednotlivých druzích stromů, tj. které z nich jsou jehličnany – *smrk, jedle, borovice,* které z nich listnáče – *buk, dub, javor, jasan, lípa,* a které jsou třeba ovocné apod. Tyto informace ve slovnících obvykle jsou, ale vydolovat je by se mohl pokoušet jen opravdu velmi zarputilý uživatel. Prototypické definice ukazují vždy **směrem nahoru** k **nadřazeným pojmům**, ale nikdy do strany **k výrazům stejného typu, sourozencům** (coordinates) nebo směrem dolů **k hyponymům**. Každý z nás zná spoustu věcí o stromech, které by lexikografové nezačlenili do definice: víme, že stromy mají kůru, rostou ze semen, poskytují stín a chrání před větrem, rostou volně v lesích, jejich dřevo slouží jako stavební materiál nebo palivo, energii pro svůj

růst získávají fotosyntézou. Lexikografové uvádějí v definicích jen **důležité distinkce**, pouze připomínají uživateli něco, o čem se předpokládá, že to už zná, a nenabízejí mu souhrn encyklopedických znalostí. Poznamenejme tedy závěrem, že velká část těchto chybějících informací má spíše **strukturní** než faktuální povahu a že konvenční slovníky ani tak nestrádají nedostatkem informací, problémem je hlavně jejich organizace, která díky abecednímu uspořádání hesel odděluje od sebe spolehlivě věci, které by bylo užitečné mít pohromadě.

V poslední době se věnuje značná pozornost lexikální sémantice s cílem vytvořit lexikální zdroje, které by popisovaly významy lexikálních jednotek a jejich vztahy formálně (algoritmicky) a díky tomu umožňovaly i systematické využívání v oblasti počítačového zpracování přirozeného jazyka (NLP).V jednom směru začaly vznikat tzv. **strojově čitelné slovníky** (Machine Readable Dictionaries) a práce na nich ukázaly, že dosavadní standardní slovníky trpí mnoha nekonzistencemi, z nichž uveďme aspoň jednu typickou: užití odlišných hyperonym v definicích tam, kde by bylo vhodné pracovat jen s jedním. Např. v SSČ (1994) nacházíme rozdílné definice u hesel ***stůl:*** *kus nábytku tvořený vodorovnou deskou ...,* ***židle:*** *lehce přenosný kus nábytku (s opěradlem)...,* ***křeslo:*** *pohodlné sedadlo s opěradly ...*), ačkoliv je zjevné, že křeslo je také kusem nábytku.

Poznamenejme, že pro češtinu žádný strojově čitelný slovník fakticky nemáme: současná elektronická verze SSČ na CD ROM (Leda, 1998) neprošla žádnými úpravami, které by vedly ke zkonzistentnění způsobu popisu významů lexikálních jednotek a k formalizovanější organizaci struktury hesel, ani není vybavena lepšími technikami vyhledávání, takže představuje právě jen pouhý počítačový „obraceč stránek".

Dalším směrem, který se v poslední době prosazuje, je budování počítačových lexikálních databází či vytváření elektronických verzí již existujících thesaurů – zejména Rogetova, (Chapmanova revidovaná verze, 1977), dále vznik sémantických sítí WordNet (Miller et al., 1993) a EuroWordNet (Vossen et al.,1999) a systémů jako CyC (Lenat and Guha, 1990), ACQUILEX (Briscoe, 1991) a COMLEX (Grishman, Macleod, Myers, 1994).

2. Lexikální databáze jako sémantická síť – WordNet

V dalším se budeme věnovat prvním dvěma zmíněným výše, tj. lexikálním databázím: WordNetu, který již dospěl do verze 1.6 a je dílem G.A.Millera a jeho skupiny z Princetonu (viz též ftp server *clarity.princeton.edu*), a EuroWordNetu, jenž vznikl v Evropě. Za zmínku stojí, že G. A. Miller byl zpočátku blízkým spolupracovníkem N. Chomského a podílel se s ním na dvou fundamentálních kapitolách v příručce Handbook of Mathematical Psychology, (Introduction to Formal Description of Natural Language, Finitary Models of Language Users) publikované v r. 1967 (Chomsky, Miller, 1967). Zatímco Chomsky se více méně stále přidržuje svých názorů na primárnost syntaktické roviny v popisu jazyka, G. A. Miller obrátil plně svou pozornost k lexikální sémantice a jako psycholog a psycholingvista se pokusil o přístup, který charakterizuje jako **psycholexikologii.** V jejím rámci usiluje spolu s Johnsonem-

-Lairdem (Miller, Johnson-Laird, 1976) o poznání toho, jak je organizována naše lexikální paměť, na jakých principech jsou budovány naše mentální slovníky. Počátek psycholexikologie je spojen se studiem slovních asociací, s pokusy o modelování mentálního slovníku, výchozí myšlenkou bylo organizovat slovník konceptuálně spíše než abecedně. Tento výzkum ho přivedl k pokusu vytvořit právě WordNet.

2. 1 Psycholingvistické předpoklady

Většina psycholingvistů se shoduje v tom, že anglická obecná substantiva jsou v sémantické paměti organizována hierarchicky, ale není definitivně jasné, zda generické informace se dědí nebo jsou jen redundantně uloženy. Quillian (1968) – byl první, kdo to formuloval explicitně. Experimentální asociační testy Collinse a Quilliana (1969) vycházely z předpokladu, že reakční časy mohou vypovídat o počtu hierarchických rovin oddělujících dva významy. Pokusy ukázaly, že reakční čas při odpovědi PRAVDA na podnět „*A canary can sing.*" je kratší než při odpovědi PRAVDA na podnět „*A canary can fly.*" Reakční doba na podnět „*A canary has skin.*" je ještě delší. Interpretace je taková, že „*can sing*" je v sémantické paměti uloženo jako příznak „*canary*", „*can fly*" jako příznak „*bird*" a „*has skin*" jako příznak „*animal*". Kdyby všechny tři rysy byly uloženy jako příznaky „*canary*", měly by všechny být akceptovány stejně rychle. Reakční časy nejsou stejné, protože „*can fly*" a „*has skin*" se patrně zpracovávají jako nadřazené, což trvá déle. Collins a Quillian z toho vyvodili závěr, že generická informace není uložena redundantně, ale zpracovává se, když je to potřeba.

Psychologická evidence, že znalost významů substantiv je v sémantické paměti organizována hierarchicky, se opírá rovněž o další poznatek, že lidé velmi snadno zpracovávají anaforické výrazy a komparativní konstrukce. Lze říci: že (a) nadřazená substantiva mohou sloužit jako anaforické výrazy odkazující zpět ke svým hyponymům. Pak v konstrukci

(k1) *Vlastnil pušku, ale z té zbraně se nikdy nevystřelilo.*

se *zbraň* bezprostředně chápe jako anaforický výraz s antecedentem *puška*.

Dále: (b) nadřazené výrazy a jejich hyponyma se nedají dost dobře srovnávat (Bever, Rosenbaum, 1970). Konstrukce

(k2a) *Puška je bezpečnější než zbraň.*,

případně

(k2b) *Zbraň je bezpečnější než puška.*

nejsou, jak lze vidět, sémanticky docela v pořádku.

2.2 Struktura WordNetu

WordNet čili slovní síť je slovník podle autorů založený na psycholingvistických principech. Např. ve verzi 1.5 obsahuje téměř 120 000 hesel – z toho cca 67 000 jednoduchých slovních tvarů a kolem 53 000 kolokací. To dává přes 91 000 slovních významů či synonymických řad (synsets). Nejvýraznější rozdíl mezi WordNetem a standardními slovníky je mj. v tom, že WordNet člení slovník do pěti kategorií: substantiva, verba, adjektiva, adverbia a funkční slova

(synsémantika). Fakticky jsou synsémantika ponechána stranou, to se opírá o pozorované řečové projevy afatických pacientů, z nichž vyplývá, že funkční slova jsou s velkou pravděpodobností uložena odděleně od ostatní slovní zásoby a tvoří součást syntaktické složky jazyka.

Uvedené členění se opírá o asociační experimenty, které ukazují, že když informanti měli reagovat prvním slovem, které je napadlo, na předložená slova patřící k různým slovním druhům, reakce vypadaly následovně:

na substantiva – substantivem : 79 %
na adjektiva – adjektivem : 65 %
na slovesa – slovesem : 43 %.

Dále se WordNet liší od standardních slovníků v tom, že jednotlivé slovní druhy jsou v něm organizovány rozdílně – přihlíží se důsledně k jejich odlišné sémantické povaze:

— substantiva jsou ve WordNetu (modelu lexikální paměti) organizována jako **tématické hierarchie,**
— slovesa jsou organizována na základě různých vztahů **vyplývání** (entailment),
— adjektiva a adverbia jsou organizována jako **n-dimenzionální hyperprostory** *(množiny n-tic).*

Každá z těchto struktur reflektuje různý způsob organizování lexikální zkušenosti – pokusy nakládat jediný organizační princip na všechny syntaktické kategorie by znamenaly chybnou reprezentaci psychologické komplexnosti lexikální znalosti.

Výrazným rysem WordNetu je též pokus organizovat lexikální informace v termínech slovních významů, a nikoli slovních tvarů. V tomto ohledu se WordNet blíží více thesaurům než standardním slovníkům (viz např. Roget's International Thesaurus, 1977).

Výchozím bodem pro lexikální sémantiku ve WordNetu je **zobrazení** mezi formami a významy, jinak řečeno, mezi lexikalizovanými koncepty a formami, které je vyjadřují. Vychází se z předpokladu, že **různým syntaktickým kategoriím slov (slovním druhům) odpovídají různé druhy zobrazení.** Přiřazení forem a významů je víceznačné, tj. některým formám odpovídá více různých významů a některé významy mohou být vyjádřeny několika různými formami. Polysémii a synonymii lze pak chápat jako **komplementární aspekty** tohoto zobrazení, posluchač nebo čtenář rozpoznávající nějakou formu se musí vyrovnat s její polysémií, mluvčí nebo pisatel usilující o vyjádření významu se musí rozhodovat mezi synonymy.

Lexikální paměť lze tedy chápat jako organizovanou **stromově** (což umožňuje vyhnout se cirkularitám a smyčkám), kde základním vztahem ve stromové struktuře je transitivní a antisymetrický významový **vztah ISA** (*is a kind of, je druhu*) nebo jinými slovy vztah **hypero/hyponymie** vedoucí od specifického ke generickému, tj. vztah **generalizace,** k němuž opakem je vztah **specializace.** Substantiva mají obvykle jedno hyperonymum a řadu hyponym, která se ve standardních slovnících zpravidla neuvádějí. Proto je vhodné navrhnout lexikální databázi tak, že v ní jsou zakódovány oba vztahy, jak vztah generalizace, tak

i vztah specializace. Výsledkem pak je lexikální databáze typu WordNet, která se vyznačuje hierarchickou strukturou a umožňuje prohledávání shora dolů i zdola nahoru stejnou rychlostí. Uvedený princip je dobře znám v oblasti informačních technologií, kde se mluví o **systémech s dědičností** (Touretzky, 1986).

2.2.1 Sémantické vztahy ve WordNetu

Jak jsme už naznačili, ve WordNetu se pracuje s následujícími sémantickými vztahy:

— **hyponymie/hyperonymie,** který je chápán jako vztah významové podřazenosti a/nebo nadřazenosti (ISA-vztah). Je tranzitivní a antisymetrický a generuje hierarchickou (stromovou) reprezentaci pro substantiva.

— **synonymie** – je ve WordNetu nejzávažnějším vztahem: nevysvětluje sice, co jednotlivé významy jsou, ale vyznačuje, že existují a liší se od sebe. V podstatě je tu synonymie chápána v duchu Leibnizovy definice založené na pojmu substituovatelnosti, ale oslabené o vztažení ke kontextu. Výrazy spojené vztahem synonymie se seskupují do **synonymických řad** (synsets), které jsou základním organizačním prvkem sémantické sítě. Vztah synonymie si také vynucuje oddělení jednotlivých slovních druhů ve WordNetu, protože lexikální jednotky patřící k různým syntaktickým kategoriím nelze volně substituovat. To je v souladu s psycholingvistickou evidencí, která ukazuje, že jednotlivé slovní druhy jsou v sémantické paměti organizovány nezávisle.

— **antonymie** – je zdánlivě jednoduchý symetrický vztah, který, jak se ukazuje, není snadné přesně charakterizovat díky jeho poměrně značné komplexnosti, i když uživatelé jazyka s ním potíží nemívají. Je centrálním organizujícím vztahem pro adjektiva a adverbia.

— **meronymie/holonymie,** jenž lze charakterizovat jako vztah část – celek. Je v zásadě tranzitivní a antisymetrický a rovněž vede k budování hierarchických struktur.

2.2.2 Hyponymie/hyperonymie – substantiva a lexikální dědičný systém

Popis významu substantivních synsetů (celkem 60 000) je ve WordNetu (obvykle) založen na nadřazeném výrazu (termu) doplněném o rozlišující příznaky. Vztah hypero/hyponymie generuje hierarchickou sémantickou strukturu (má formálně podobu grafu-stromu), v níž synsety (synonymické řady) jsou propojeny ohodnocenými ukazateli (pointry). Hierarchie mají omezenou hloubku, zřídka přesahují 12 úrovní. Rozlišující příznaky jsou zavedeny tak, že tvoří lexikální systém s děděním, tj. systém, v němž každé slovo dědí všechny rozlišující příznaky všech svých nadřazených výrazů. Pracuje se také s antonymií, ale ta se u substantiv nepokládá se fundamentální organizační princip. V původní verzi se rozlišovalo 25 tematických souborů a každý z nich byl spojen s jednou primitivní sémantickou složkou. Těchto 25 hlavních hyperonym ve WN 1.5 pak fungovalo jako generické koncepty, z nichž vycházejí jednotlivé hierarchie (sémantická pole). Díky tomu, že všechny příznaky, které charakterizují jednotlivé počátky, se dědí na všechna hyponyma, lze jednotlivé začátky hierarchicky strukturovaných sémantických polí pokládat za **primitivní sémantické**

příznaky všech slov v daném poli. To je dobře vidět v Tab.1, která obsahuje zmíněných 25 původních počátků – většina substantiv ve WordNetu 1.5 spadá právě pod ně. Zajímavé je, že uvedená sémantická pole jsou celkem mělká, zřídka hlubší než 10 úrovní, lidské výrobky jako dopravní prostředky mívají kolem 7–8 úrovní, např.: *sedan – vůz – motorové vozidlo – kolové vozidlo – dopravní prostředek – lidský výtvor – věc.* Lidské hierarchie mívají kolem 3–4 úrovní.

Tab.1 Vrcholová hyperonyma ve WordNetu 1.5

act, action, activity (činnost, aktivita)	natural object (fyzický objekt)
animal, fauna (zvíře, fauna)	natural phenomenon (přírodní jev)
artefakt (výtvor, výrobek)	person, human being (osoba, lidská bytost)
attribute, property (atribut, vlastnost)	plant, flora (rostlina, flora)
body, corpus (tělo, těleso)	possession (vlastnictví)
cognition, knowledge (znalost, poznání)	process (proces)
communication (komunikace, sdělování)	quantity, amount (kvantita, množství)
event, happenning (událost)	relation (vztah)
feeling, emotion (pocit, emoce)	shape (podoba, tvar)
food (potrava, jídlo)	state, condition (stav)
group, collection (skupina, soubor)	substance (substance, látka)
location, place (umístění, místo)	time (čas)
motive (motiv)	

Těchto 25 počátků odpovídá potom v EuroWordNetu položkám tvořícím vrcholovou ontologii, jichž je však o něco více – 63 (viz níže).

2.2.1.2 Adjektiva – atributy a modifikace

Celkem je ve WordNetu cca 16 000 adjektivních synsetů, které se člení na dvě rozsáhlé třídy: **deskriptivní** a **relační**. První připisují (obvykle) svým řídicím substantivům hodnoty bipolárních atributů a jsou tedy organizována v termínech binárních opozic antonymních *(velký: malý)* a podobných významů (synonym).

K relačním adjektivům patří adjektiva jako *prezidentský, nukleární, zubní,* mají tedy vztah k určitému substantivu nebo jsou s ním nějak spojena, nerozlišují škály a neodkazují k vlastnosti svého řídicího substantiva, nemají přímá antonyma a nelze je stupňovat. Ve WordNetu je jich kolem 1700.

Samostatně stojí malá a uzavřená skupina referenčně modifikujících adjektiv jako *předchozí* nebo *údajný.* Samostatnou skupinu představují také adjektiva označující barvy.

2.2.1.3 Slovesa

Ve WordNetu je nyní něco přes 11 000 slovesných synsetů. Díky své významové flexibilitě se slovesa obecně vyznačují vyšší polysémií – např. Collinsův slovník (1990) uvádí u substantiv 1,74 významu na substantivum, u sloves to

činí v průměru 2,11. Sémanticky se slovesa podstatně liší od ostatních slovních druhů svou predikátově argumentovou strukturou a vazbami na své aktanty, proto nejsou organizována na základě vztahu hypero/hyponymie, nýbrž na základě vztahu vyplývání (*prodávat : platit*) a jeho modifikací: troponymie (*chrápat : spát*) a kauzálních vztahů (*dát : mít*). Rozlišuje se 15 hlavních slovesných významových tříd (Levin, 1989), konkrétně slovesa tělesných funkcí, změny, poznání, komunikace, soutěžení, spotřeby, kontaktu, tvoření, emocí, pohybu, vnímání, vlastnění, sociální interakce a slovesa označující počasí.

3. Lexikální databáze EuroWordNet-1 a 2

WordNet 1.5 vytvořený G. A. Millerem a jeho skupinou pokrývá dostatečně (americkou) angličtinu a díky svým vlastnostem se stal impulsem pro podobné aktivity v Evropě. V r. 1997 se skupina lexikografů kolem P. Vossena z university v Amsterdamu rozhodla začít budovat sítě slov pro tři vybrané západoevropské jazyky, a to v rámci projektu EuroWordNet-1. Na ten pak v r. 1998 navázal EuroWordNet-2, do něhož byly zahrnuty další čtyři jazyky, z toho dva východoevropské.

3.1 EuroWordNet 1 – angličtina, holandština, italština, španělština

Projekt EuroWordNet (dále EWN) jako celek vychází z princetonského WordNetu 1.5 a jeho hlavním cílem bylo nejprve rozšířit budování sítě slov na tři evropské jazyky, tj. holandštinu, italštinu a španělštinu, a posléze na další čtyři – němčinu, francouzštinu, češtinu a estonštinu. Nově budované slovní sítě rovněž obsahují informace o substantivech, slovesech, adjektivech a adverbiích a opírají se o pojem synonymické řady (synsetu). Připomeňme, že každý synset zahrnuje jeden nebo více významů slov, které lze pokládat za významově totožné nebo blízké, spolu s glosou definující daný význam. Jako příklad uveďme synset pro lexikální jednotku *soubor:*

soubor:2, datový soubor:1 – (množina záznamů vztahujících se k sobě a ukládaných pohromadě)

Synset je tedy tvořen posloupností *soubor:2, datový soubor:1*, tj. *soubor* ve významu 2 je
synonymní s výrazem *datový soubor* ve významu 1.

Synsety mohou vstupovat do předem definovaných sémantických vztahů (0 nebo více), jako jsou **hyponymie, hyperonymie, meronymie** a **holonymie** a další. Daný synset může mít u sebe uveden vztah ke svým:

antonymům (*dobrý : zlý*)
hyperonymům (*auto : dopravní prostředek*)
hyponymům (*pták : kanárek*)
meronymům (*dveře : zámek*)
holonymům (*ruka : tělo*)
sourozencům (*pes : vlk : kojot : hyena*)
vyplývajícím výrazům (*kupovat : platit*)
kauzacím (*rozbít : rozpadnout se*)

V rámci projektu EuroWordNet se tedy nejprve budovala lexikální databáze EWN-1, která vedle WordNetu 1.5 (tj.angličtiny) zahrnovala i holandský, španělský a italský wordnet. Proti WordNetu 1.5 byly provedeny některé úpravy a změny, které spočívají v zavedení:

- **vrcholové ontologie (top ontology – TO),** která je chápána jako hierarchie jazykově nezávislých konceptů a odráží význačné sémantické distinkce, např. *předmět* a *substance*, *dynamický* a *statický*. Zahrnuje celkem 63 základních sémantických komponent vybraných s přihlédnutím k různým sémantickým teoriím a paradigmatům. Výchozí rámcovou představu o konstruktech ve vrcholové ontologii poskytuje Tab.1 výše.
- **množiny základních konceptů (base concepts – BC)** tvořené 1000 základními koncepty, které jsou vybrány na základě obecně sdíleného sémantického rámce, jímž je vrcholová ontologie. Základní koncepty reprezentují sdílená jádra jednotlivých sítí slov, na druhé straně se také od sebe liší v závislosti na povaze jednotlivých začleněných jazyků. Představují nejdůležitější významy převažující v jednotlivých lokálních wordnetech a tvoří jádro multilinguální databáze. Proto jsou také propojeny prostřednictvím vrcholové ontologie navržené speciálně k tomuto účelu. Aby se dosáhlo maximální shody, wordnety se budují shora dolů tak, že se začíná právě množinou základních konceptů zvolených na základě společného sémantického rámce.
- **jazykově nezávislého souboru indexů (interlingual index – ILI),** který představuje hlavní novum ve vztahu k výchozímu WordNetu 1.5. ILI tvoří nestrukturovaný seznam významů, kde každý ILI-záznam se skládá ze synsetu a glosy a specifikuje význam a odkaz ke svému zdroji. Mezi jednotlivými ILI-záznamy jako takovými se neudržují žádné vztahy. Budování úplné jazykově neutrální ontologie se pokládá za příliš komplexní a časově náročné vzhledem k časovým omezením projektu. Hlavní výhodou tohoto designu je, že jazykově specifické vztahy a vztah ekvivalence se nemusí uvažovat z hlediska více-víceznačného zobrazení mezi jednotlivými jazyky vstupujícími do databáze EuroWordNet.
- **vztahů ekvivalence (EQ-relations)** – ty jsou zavedeny mezi ILI a jednotlivými sítěmi slov a umožňují vztahovat k sobě a porovnávat jednotlivé wordnety. Pomocí vhodných nástrojů (viz níže o Polarisu) lze pak automaticky vytvářet projekce z jedné sítě slov do druhé.

3. 2 EuroWordNet-2 – francouzština, němčina, čeština, estonština

V návaznosti na EWN-1 hlavními cíli projektu EuroWordNet-2 (Vossen et al, 1998) jsou:

- Definice obecné množiny základních konceptů (BC) pro všechny jazyky EWN-1 a EWN-2: je to soubor významů, jež hrají klíčovou roli v jednotlivých wordnetech. Stanovený rozsah = 1000 synsetů, z toho je 700 substantivních a 300 verbálních.
- Zachycení vnitřně jazykových vztahů (ILR) a vztahů ekvivalence v rámci základních konceptů (BC) pro němčinu, francouzštinu, estonštinu a češtinu. Výsledkem budou jádra wordnetů, každé v rozsahu 7 500 synsetů, z toho je

5 000 substantivních a 2 500 synsetů. Adjektiva a adverbia zatím zůstávají stranou, ale s jejich zpracováním se počítá.

- Průběžná aktualizace jazykově nezávislého souboru indexů (ILI) o další významy, které je potřeba doplnit pro potřeby toho kterého jazyka a které nebyly v původním Wordnetu 1.5 obsaženy. Tím se dosáhne i lepší shody mezi jednotlivými sítěmi slov.
- Integrace jednotlivých wordnetů do společné databáze EuroWordNet 2, jejich porovnání a ověření vzájemné kompatibility.

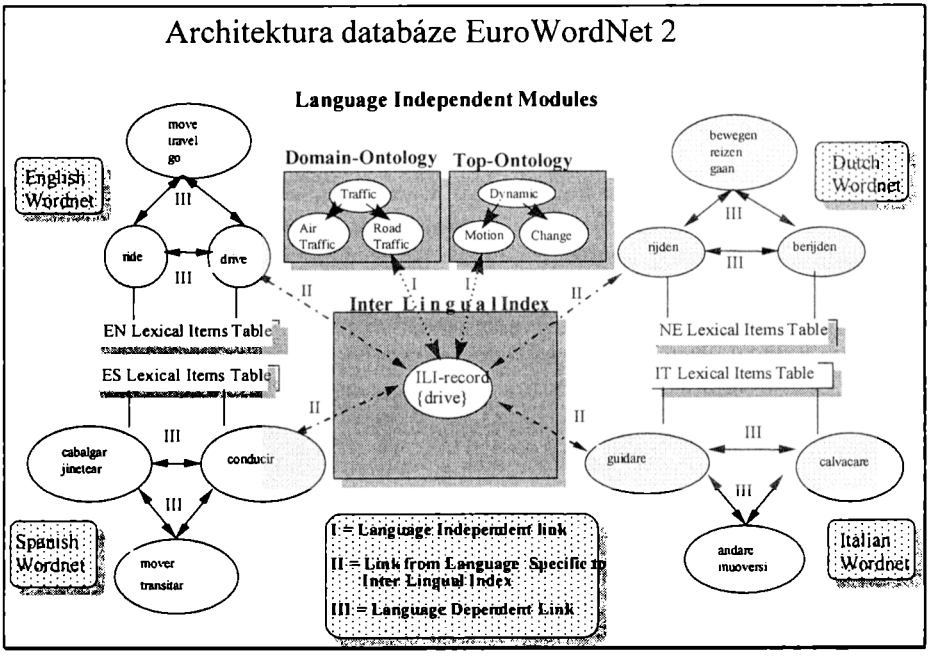
Můžeme tedy shrnout hlavní body, v nichž se EWN odlišuje od Wordnetu 1.5. Jsou to:

a) **multilingualita** databáze EuroWordNet 2 – je jí dosaženo tím, že se rozlišuje mezi jazykově specifickými moduly a odděleným jazykově nezávislým modulem (ILI). Každý z jazykových modulů reprezentuje jedinečný jazykově specifický systém vnitřních jazykových vztahů mezi synsety. Každý synset rovněž obsahuje vztah ekvivalence k synsetu v jazykově nezávislém souboru indexů (ILI). ILI-synset neboli ILI-záznam je částí jazykově nezávislého modulu a může být označen jako patřící do nějaké domény nebo mající vztah k nějakému jazykově nezávislému vrcholovému konceptu. Vrcholové koncepty reprezentují fundamentální sémantické distinkce jako např. *předmět : substance* nebo *životnost : neživotnost* a další. Synsety tvořící ILI jsou převážně odvozeny z WordNetu 1.5, ale budou rozšířeny použitím speciálního aktualizačního programu v případě, že specifické významy z jiných jazyků nejsou ve WordNetu 1.5 přítomny a vyžadují to. Konečný ILI tak bude **nadmnožinou** všech konceptů vyskytujících se v různých wordnetech. Skrze ILI lze mít přístup k dalším wordnetům tak, abychom našli synsety napojené na stejné synsety a verifikovali způsob, jak se k sobě vzájemně vztahují. Bylo navrženo speciální multilinguální rozhraní, které umožní srovnávat vztahy ekvivalence a struktury sémantických polí napříč jednotlivými wordnety.

b) Dalším rozdílem je to, že u lexikální databáze EuroWordNet-2 se již počítá se systematickým využitím v oblasti strojového zpracování informací (Information Retrieval), konkrétně s multilinguálními aplikacemi pro internetové prohlížeče a pro lexikální zdroje použitelné v systémech strojového překladu nové generace. Dále se počítá s dosažením maximální kompatibility vzhledem k různým zdrojům a současně i s tím, že ve wordnetech se zachovají vztahy specifické pro jednotlivé jazyky.

Na obr. 1, který ukazuje základní strukturu databáze EUWN 2, lze vidět vrcholový koncept *Motion (pohyb)*, který je v tomto případě bezprostředně napojen na ILI-záznam *drive* (*řídit*) a díky tomu se nepřímo vztahuje také na všechny jazykově specifické koncepty spojené s tímto ILI-záznamem. Prostřednictvím vnitřně jazykových vztahů lze daný vrcholový koncept dále dědit na všechny další napojené jazykově specifické koncepty. Tak lze budovat jednotlivé wordnety na základě společného rámce, v němž se lexikalizace seskupené kolem daných základních konceptů mohou od jazyka k jazyku lišit. Ve schématu se také

objevuje doménová hierarchie, která obsahuje znalostní struktury, jež seskupují významy v termínech témat nebo scénářů, např. sem patří silniční doprava, vzdušná doprava, sporty, nemocnice, restaurace apod., v rámci EWN-1,2 však zatím není implementována;



Obr.1 Architektura databáze EuroWordNet 2

4. Budování české slovní sítě – českého WordNetu, dosavadní výsledky

Zatím je k dispozici český WordNet v rozsahu cca 8000 synsetů (asi 1200 slovesných, zbytek – 6 800 substantivních. Při jeho vytváření bylo použito následujících zdrojů:

a) *Výkladový slovník češtiny*, což je pracovní název postupně budované lexikální databáze češtiny, která má dnes přibližně 55 000 hesel a 65 000 významů. Od např. SSČ se podstatně liší v tom, že je systematicky budována jako důsledně formalizovaná textová databáze (na principech podobných SGML) a s důrazem na maximální vnitřní konzistenci.

b) *Lingea Lexicon 2.0 angličtina* (Lingea s.r.o, 1998), což je oboustranný elektronický A-Č a Č-A slovník, který v současné podobě obsahuje ve směru Č-A asi 54 000 hesel a 58 000 významů a ve směru A-Č zhruba 78 000 hesel a 102 400 významů. Toto dílo mimo jiné zahrnuje i automatické morfologické slovníky angličtiny i češtiny a jádro programu LEMMA (Ševeček, 1996), díky nimž rozpoznává libovolné české i anglické tvary slov.

c) *Slovník českých synonym*, (Pala -Všianský, 1994), obsahující v aktuální verzi přibližně 20 000 hesel a 15 000 synonymických řad (synsetů), jichž bude po

potřebných úpravách použito pro synsety začleněné do české sítě slov. Existuje v elektronické verzi a rovněž funguje s automatickou lemmatizací.

Pomocnými lexikálními zdroji jsou dále:

d) *Seznam českých kolokací* obsahující nyní asi 18 000 položek, byl získán z textového korpusu *ESO* (viz níže), který je budován a udržován na Fakultě informatiky MU. Seznam kolokací byl získán statistickými technikami – výpočtem parametru vzájemné informace (Pala, Rychlý, 1998), a je dále tříděn podle četností a dalších syntaktických kritérií – slovosledu a slovních druhů. Seznam kolokací bude v blízké budoucnosti doplněn a rozšířen, jakmile budou spočítány parametry vzájemné informace (MI score) i pro aktuální verzi Českého národního korpusu.
e) Gramaticky i strukturálně značkovaný korpus *DESAM* (Pala, Rychlý, Smrž, 1998), který vznikl na Fakultě informatiky Masarykovy university v průběhu posledních dvou let jako součást Českého národního korpusu. Jeho rozsah je něco přes 1 mil. českých slovních tvarů.
f) Textový korpus *ESO* budovaný na Fakultě informatiky v průběhu r. 1998 z novinových a publicistických textů (1996–98), jeho aktuální rozsah činí 61 mil. českých slovních tvarů, jedna jeho verze je částečně lemmatizována.

5. Nástroje

Je zjevné, že popisovanou síť slov lze sotva budovat manuálně, má-li vzniknout v rozumném časovém úseku a s přijatelnými náklady. Při sestavování české sítě se tedy neobejdeme bez použití počítačů a vhodného softwaru, který se vyvíjí v průběhu budování databáze. Při vytváření českého wordnetu se nyní používají následující programové nástroje:

1) **Polaris** – specializovaný program založený na technologii FLAIM firmy Novell. Je uzpůsoben pro potřeby projektu EuroWordnet 2, umožňuje jednotným způsobem prohlížet současně sítě slov všech zúčastněných jazyků. Zobrazuje ve formě stromu hyperonyma i hyponyma zvoleného synsetu, v případě hyponym lze zobrazit buď nejbližší následníky, nebo tranzitivně všechna hyponyma. Také je možno provádět projekci vybrané množiny synsetů do jiného jazyka a tak konfrontovat zastoupení jednotlivých sémantických polí v různých jazycích. Program dále umožňuje importovat synsety z přesně definovaného textového formátu, případně exportovat zvolené části databáze do textové podoby.

2) **EWN-tools** je sada konverzních programů a filtrů umožňující dávkové zpracování dat českého wordnetu. V zásadě dovolují následující:

a) konverzi mezi externím textovým formátem programu Polaris a vlastním textovým (databázovým) formátem umožňující efektivnější dávkovou i editační práci s daty,
b) automatické doplnění možných českých ekvivalentů k vybraným synsetům Wordnetu 1.5,
c) automatické doplnění vztahů ekvivalence v těch případech, kdy uvedený literál anglického slova (resp. anglických slov) toto určuje jednoznačně,

d) automatické doplňování ILI-indexů podle symbolického označení vztahu ekvivalence libovolným prvkem synsetu,
e) automatické vytváření synsetů českého wordnetu na základě shodnosti ILI-indexů,
f) třídění synsetů podle slovních druhů a některých dalších gramatických kategorií a opětovné slučování a zatřiďování hesel a synsetů.

3) **Lingea Lexicon** – program pro efektivní prohlížení anglicko-českého a česko-anglického slovníku firmy Lingea byl doplněn o možnost zobrazování hesel slovníku Wordnet 1.5 včetně všech vnitřně jazykových vztahů, zvláště pak hyperonym a hyponym. Dále umožňuje stejným způsobem prohlížet i český slovník synonym uvedený výše. Lexicon spolu s programem Polaris tvoří základní pomůcky pro interaktivní rozšiřování a zpřesňování databáze české sítě slov.

4) **Lemmatizátor** – nezbytnou pomůckou při práci je i český a anglický lemmatizátor s názvem LEMMA (Ševeček, 1996). Ten byl použit a používá se např. při zjišťování vhodných kandidátů pro české základní koncepty, pro značkování korpusu ESO (viz výše), ze kterého se získávají frekvenční informace o zastoupení jednotlivých hesel v současné češtině nebo informace pro výpočet pravděpodobnosti souvýskytu určitých hesel, tj. parametru tzv. vzájemné informace (Pala, Rychlý, 1998). Pomocí obrácené funkce lemmatizátoru, tj. generování tvarů, lze rovněž zrekonstruovat základní podobu potenciálních českých kolokací.

6. Závěry

Integrace národních wordnetů vznikajících v rámci EuroWordNetu-1,2 zajistí maximální kompatibilitu mezi wordnety jednotlivých jazyků a umožní opravdovou multilingualitu připojením dalších zdrojů do sdílené databáze EuroWordNet. Rozšíření také posílí roli technologie EuroWordNetu a jeho datových formátů jako de facto standardu pro reprezentaci lexikálně sémantických dat v rámci evropské informační společnosti. Takový standard nejen umožní budoucí inkorporaci dalších jazyků, ale také poskytne jedinečné rozhraní pro lexikálně sémantická data softwarovým vývojářům v informačním průmyslu. V delším časovém úseku se wordnety pravděpodobně mohou stát **páteří** jakýchkoli sémantických databází a v blízké budoucnosti otevřou dveře celé řadě nových aplikací a služeb v Evropě na mezinárodní i transkulturní úrovni. Rozsah lexikální databáze je takový, aby byla kompatibilní se slovníky Parole (Parole, 1998). V kombinaci s těmito slovníky bude poskytovat základ pro budování kvalitních jazykových technologií pro hlavní evropské jazyky včetně uvedených východoevropských.

Celkově vzato, rozšíření EWN-2 lze tak částečně chápat jako integrační a standardizační úlohu, která zvětšuje rozsah a impakt EWN-1 a buduje cestu k vytvoření sémantických zdrojů pro země chystající se vstoupit do EU.

Český příspěvek k EuroWordNetu je lingvisticky významný v tom, že začleňuje do uvedeného výzkumného paradigmatu první slovanský jazyk se všemi jeho typologickými odlišnostmi a dává možnost systematického porovnání lexikálních zdrojů češtiny s hlavními představiteli románských a germánských ja-

zyků. Navíc tento projekt poskytuje příležitost srovnávat slovní zásobu češtiny i s estonštinou, jež je představitelem typologicky odlišné skupiny ugrofinských jazyků. Zkušenosti získané při budování české sítě slov už teď ukazují, které jevy v češtině nemají v ostatních jazycích přímou obdobu a způsobují při začleňování českého wordnetu do celkového rámce jisté komplikace: pochopitelně k nim patří na prvním místě slovesné vidy, deminutiva, augmentativa a také odvozeniny vznikající přechylováním a dvojitou prefixací u sloves, tj. jevy, pro které v konfrontovaných západoevropských jazycích nenacházíme odpovídající (pravidelné) lexikalizace.

Účast na projektu EuroWordNet-1,2 je rovněž vhodnou příležitostí pro konfrontaci metodologických postupů využívaných v komputační lexikografii a počítačové lingvistice u nás a na západoevropských pracovištích. S uspokojením lze konstatovat, že v dané oblasti nejsme nikterak pozadu, naopak, např. naše postupy používané při zpracování morfologie jsou díky syntetické povaze češtiny propracovány důkladněji a úplněji.

Poznámka:

Budování české lexikální databáze typu WordNet je začleněno do projektu Evropské Unie EuroWordNet-2 (LE4–8328) a je také podporováno v rámci národního výzkumného projektu VS97028 (Laboratoř zpracování přirozeného jazyka na Fakultě informatiky MU).

LITERATURA

BEVER, T.G., ROSENBAUM, P. S.: Some Lexical Structures and Their Empirical Validity, in *Jacobs and Rosenbaum (eds.), Readings in English Transformational Grammar*, Waltham, Mass, 1970.

BRISCOE, E.: Lexical Issues in Natural Language Processing, in *Klein and Veltman (eds.), Natural Language and Speech, Proceedings of the Symposium on Natural Language and Speech*, 39–68, Springer-Verlag, Berlin, 1991.

COLLINS, A. M., QUILLIAN, M. R.: Retrieval Time from Semantic Memory, *Journal of Verbal Behavior and Verbal Learning* 8, 1969, 240–247.

CHOMSKY, N., MILLER, G. A.: Handbook of Mathematical Psychology, kap. 11, 12, 13, 1967.

GRISHMAN R., MACLEOD, C., MYERS, A.: COMLEX syntax: Building a Computational Lexicon, *Proceedings of Coling'94*, 1994.

IDE, N., VÉRONIS, J.: 1998 Introduction to the Special Issue on Word Sense Disambiguation: The State of Art, *Computational Linguistics*, March 98, vol. 24, No.1

LENAT, D. B., GUHA, R. V.: Building Large Knowledge-based Systems, Addison Wesley, 1990.

LEVIN, B.: Towards a Lexical Organization of English Verbs, Ms., Evanston: Northwestern University, 1989.

MILLER, G. A. et al.: Five Papers on WN, 1990, revised version, August 1993.

New Oxford Dictionary of English (NODE), Oxford University Press, Oxford, 1998

PALA, K., RYCHLÝ, P.: Mutual Information in Corpus ESO, Proceedings of TSD'98, Brno, 1998, s.49–53

PALA, K., RYCHLÝ, P., SMRŽ, P.: Annotated corpus for Czech – DESAM, Proceedings of Sofsem'97, Springer Verlag, 1997.

PALA, K., VŠIANSKÝ, J.: Slovník českých synonym, NLN, Praha 1994.

PAROLE, projekt EU orientovaný na budování jednotných lexikálních zdrojů pro hlavní evropské jazyky, Pisa, Sheffield, Amsterdam,1998.

QUILLIAN, M., R.: Semantic Memory, in *Minsky (eds.) Semantic Information Processing*, Cambridge, Mass., MIT Press, 1968.

RUIMY, N. CORAZZARI, O., GOLA, E., SPANU, A., CALZOLARI, N., ZAMPOLLI, A.: The European LE-PAROLE Project: The Italian Syntaxic Lexicon, Technical Report, Pisa, 1996.

Roget's International Thesaurus, 4th ed. by R. L. Chapman, New York, Harper and Row, 1977.

SSČ = Slovník spisovné češtiny, Academia, 2.vyd. 1994, 3. vyd. CD ROM, Leda 1998.

ŠEVEČEK, P.: LEMMA, morfologický analyzátor a lemmatizátor pro češtinu, program v jazyce C, Brno 1995–6.

ŠEVEČEK, P., et al.: Lingea Lexicon 2.0 angličtina, Lingea s.r.o., Brno 1998.

TOURETZKY, D. S.: The Mathematics of Inheritance Systems, Los Altos, Calif.,1986.

VOSSEN, P. et al.: EuroWordNet-2: Extending EuroWordNet with Other Languages, Telematics Programme, Technical Report No 1, 1998

VOSSEN, P. et al.: Introduction to EuroWordNet, v tisku 1999.

CZECH LEXICAL DATABASE OF THE WORDNET TYPE

The presented paper deals with the electronic lexical resources in the form of semantic nets suitable for natural language processing applications. It consists of the two parts: in the first one the basic issues of the WordNet approach to building the lexical resources are presented, and the main principles of psycholexicology are outlined following G. A. Miller's (1993) conception, on which the structure and organization of WordNet 1.5 is based.

In the second part of the paper we concentrate on the extension of this research and describe EuroWordNet-1,2, a multilingual lexical database (and also EU research project of the same name) involving 8 European languages, particularly (apart from English) Dutch, Spanish, Italian, German, French, Czech and Estonian. The techniques of building Czech WordNet are demonstrated together with the main resources and tools. In the end the results are given: the first version of Czech WordNet being developed and implemented under EuroWordNet database Polaris contains now slightly more than 8000 synsets, i.e. approximately 6800 noun synsets and 1200 verbal ones.

Karel Pala
Katedra informačních technologií
Fakulta informatiky Masarykovy univerzity
Botanická 68a
602 00 Brno
pala@fi.muni.cz

Pavel Ševeček
Lingea, s.r.o.
Husova 8a
602 00 Brno
pavel@fi.muni.cz