

Fakulta informatiky Masarykovy university

Počítačové zpracování přirozeného jazyka

I.K., Karel Pala

Brno, září 2000

Obsah

0.1	Předmluva	3
0.2	Úvod	5
0.2.1	Roviny analýzy jazyka	8
0.2.2	Reprezentace a porozumění	10
0.3	Data pro zpracování přirozeného jazyka – korpusy	15
0.3.1	Jak se budují korpusy?	16
0.3.2	Typy korpusů a standardizace	17
0.3.3	Budování korpusu – sběr dat	18
0.3.4	Vnitřní struktura korpusu	19
0.3.5	Korpusové nástroje	20
0.3.6	Značkování (anotování) korpusů	21
0.3.7	Značkování pro češtinu – LEMMA, AJKA	22
0.3.8	Morfologické (gramatické) značkování	22
0.3.9	Syntaktické značkování	23
0.3.10	Situace v češtině	24
0.3.11	Struktura ČNK	24
0.4	Reprezentace morfologických struktur (pro češtinu)	25
0.4.1	Přehled notace pro českou morfologii a syntax	27
0.4.2	Algoritmický popis (české) morfologie	32
0.5	Reprezentace syntaktických struktur – gramatiky	34
0.5.1	Gramatiky pro popis PJ	34
0.5.2	Gramatika jako reprezentace znalosti	35
0.5.3	Formální gramatiky	36
0.5.4	Typy gramatik	40

0.5.5	Několik slov o PROLOGu	42
0.5.6	Gramatiky v PROLOGU	42
0.5.7	Nekontextové gramatiky a DC gramatiky	42
0.5.8	Valenční rámce a jejich začlenění do formálních gramatik	44
0.5.9	Vztah mezi slovesnými významy a valencemi	49
0.5.10	Desambiguace – metody	56
0.6	Reprezentace významu	56
0.6.1	Lexikální význam – slova a slovní spojení	56
0.6.2	Významy slov a slovníky	62
0.6.3	Lexikální databáze	63
0.6.4	WordNet a sémantické sítě	63
0.6.5	Lexikální databáze EuroWordNet-1 a 2	68
0.6.6	Budování české slovní sítě – českého WordNetu, dosavadní výsledky	71
0.6.7	Nástroje	72
0.7	Sémantické reprezentace vět PJ	74
0.7.1	Formální aparát pro SR – charakteristika TIL	75
0.7.2	Formální aparát – TIL a teorie typů	78
0.7.3	Sémantická analýza výrazů PJ	78
0.7.4	Nástin algoritmu sémantické analýzy	80
0.7.5	Poznámky k sémantické roli jmenných skupin	83
0.7.6	Referenční role funkční perspektivy větné	85
0.8	Pragmatická rovina	88
0.8.1	Interní pragmatika	88
0.8.2	Externí pragmatika	89
0.9	Dialogové systémy, inference	91
0.9.1	Analýza promluvy, promluvové objekty	91
0.9.2	Anafora, anaforické vztahy	91
0.9.3	Odkazovací výrazy, rozpoznávání antecedentů	91
0.9.4	Historie promluvy a promluvový zásobník	91
0.9.5	Segmenty v promluvě	91
0.10	Závěr	91

0.1 Předmluva

Předkládaná práce představuje pokus shrnout výzkumy v oblasti počítačového zpracování češtiny, které probíhaly od počátku 70. let na katedře českého jazyka FF UJEP v Brně, pokračovaly v Ústavu českého jazyka FF MU v průběhu osmdesátých (počínaje již 1978, viz Machová, Havel, Pala, 1978) a na počátku let devadesátých. Od r. 1995 se výzkum přesunul na Fakultu informatiky a v současnosti se soustřeďuje v Laboratoři zpracování přirozeného jazyka, která vznikla na Fakultě informatiky v r. 1997. I když jsme se této problematice věnovali systematicky již dříve: první naše experimenty s automatickou syntaktickou analýzou češtiny se uskutečnily v r. 1977 v OVC VUT na počítači TESLA 200 a poté ve spolupráci s ÚVT UJEP na minipočítači PDP 11. K zásadnímu obratu ovšem došlo až v r. 1988, kdy se po překonání tehdy četných a zcela nesmyslných administrativních překážek podařilo na katedru českého jazyka FF UJEP získat osobní počítač COMMODORE PC 40 AT. Byl to dokonce první osobní počítač na celé tehdejší FF UJEP MU, a teprve díky jeho instalování jsme v našich výzkumech mohli přejít od teoretických popisů jazyka k jejich ověřování a tedy i k získávání výsledků praktické povahy a materiálově většího rozsahu.

V experimentech na osobním počítači jsme využili zkušeností získaných předtím na minipočítači PDP 11 v ÚVT UJEP a svou pozornost jsme zaměřili na popis české syntaxe s využitím PROLOGU a aparátu DC gramatik, i když naše předchozí experimenty na minipočítači PDP 11 se opíraly o programový systém WANDER (Benešovský, Šmídek, 1984). Téměř souběžně se pak začaly práce v oblasti morfologie (Osolsobě, 1988), jejichž výsledkem byl integrovaný morfologicko-syntaktický analyzátor KLARA, který po programátorské stránce realizoval S. Franc (Pala, Osolsobě, Franc, 1987). To však byl teprve začátek ÍEHĚuŘĚÍ, – v r. 1988 se nám podařilo získat elektronickou verzi glosáře Slovníku spisovného jazyka českého (SSJČ, 1960) pořízenou brněnskými křížovkáři a čítající cca 192 000 položek. Teprve tato data umožnila navrhnout a vytvořit relativně úplný algoritmický popis české morfologie obsahující v současnosti téměř 1200 ohýbacích vzorů pro substantiva, adjektiva a zájmena, číslovky, slovesa i neohebné slovní druhy (Osolsobě, 1990, 1995, Ševeček, 1995, Sedláček, 1999).

Algoritmický popis české morfologie se pak stal východiskem a podkladem pro řadu konkrétních programových produktů: automatického korektoru (Franc, dipl. práce, 1990), prvního morfologického analyzátoru XANTIPA, na něj navazujícího a vylepšeného morfologického slovníku, analyzátoru, generátoru a také lemmatizátoru LEMMA (Ševeček, 1992, 1995) a postupně připravovaného syntaktického analyzátoru (Pala, 1992). Práce na dobudování morfologické analýzy pokračovaly a vedly k vytvoření nového morfologického analyzátoru a lemmatizátoru AJKA (Sedláček, 1999), v němž je proti programu LEMMA odstraněna řada chyb ve vzorech a který díky své otevřenější koncepci umožňuje v sobě postupně integrovat řadu pravidelných slovotvorných procesů a také vazby na lexikální informace. Nyní je tedy na FI MU pro češtinu k dispozici kvalitní morfologický modul (vedle komerčně orientovaného programu Lemma, jehož autorem je P. Ševeček, viz výše,

a morfologického analyzátoru vytvořeného J. Hajičem, viz Hajič 2000), který se využívá několika způsoby: jako lemmatizátor, morfologický značkovač, a zejména jako první stupeň syntakticko-sémantického analyzátoru (Horák, Smrž, 2000, Hadacz, 2000, Žáčková, 2000). S jeho pozdějším využitím se také počítá v syntéze řeči, konkrétně v systému DEMOSTHENES a jemu podobných systémech pro převod textu na řeč (TTS) (Kopeček, Kopeček, Pala 2000). Morfologický modul AJKA obsahuje nyní cca 150 000 českých kmenů a 1200 vzorů a je dále doplňován z korpusových zdrojů.

Je tedy vcelku přirozené, že materiálově i implementačně zatím nejlépe zpracovaná část jazykového systému češtiny zahrnuje především rovinu morfologickou, zatímco podobné zvládnutí roviny syntaktické si ještě vyžádá nemalého úsilí a dalších empirických pozorování, která v současném výzkumu dosud chybí, např. tu máme na mysli širší a systematické zpracování valence českých sloves adjektiv, substantiv včetně dalších okruhů otázek. V tomto bodě se však situace výrazně mění k lepšímu: nedávno jsme dokončili výchozí valenční slovník českých sloves, který po doplnění čítá téměř 15 000 položek (Pala, Ševeček, 1996).

Vedle toho je tu i příznivá okolnost, že díky rozběhnutým se pracím na Českém národním korpusu (ČNK, buduje se v Ústavu českého národního korpusu na FF UK) je již k dispozici základní část Českého národního korpusu, čítající v současnosti cca 200 mil. českých slovních tvarů. Dalším pozitivním faktem je, že i na Fakultě informatiky vzniklo od r. 1996 několik českých korpusů – jsou zde instalovány korpusy DESAM (plně gramaticky značkováný a čítající 1 mil. slovních tvarů), korpus ESO v rozsahu 160 mil. slovních tvarů a korpus FIT obsahující texty z oblasti informačních technologií. Díky této skutečnosti se podmínky pro práci s jazykovým materiálem podstatně a příznivě mění: potřebná zkoumání mohou být spolehlivější a hlavně dostáváme možnost zjišťovat fakta, která bychom při ručním zpracování nikdy získat nemohli. Důležité je i to, že práce na korpusu a zejména na jeho značkování (tagging) jsou spojeny s budováním programových nástrojů, které se v určitém ohledu překrývají s dosavadním základním výzkumem v oblasti morfologie a syntaxe, směřují však k jedinému cíli.

U roviny sémantické jde především o nalezení co nejexpresivnějšího formálního (logického) aparátu, který by mohl sloužit jako spolehlivý nositel sémantických reprezentací vět přirozeného jazyka (češtiny). Opírajíce se o dřívější společné práce s P. Maternou a A. Svobodou, dáváme přednost aparátu transparentní intenzionální logiky (TIL, Tichý, 1989), ovšem právě zde stojí před námi ještě značná práce empirická. Její hlavní část podle našeho přesvědčení spočívá ve vytvoření vhodného sémantického slovníku, který bude moci vhodně integrovat slovníkové informace morfologické a syntaktické s logickými (o logických typech) a využívat jich v algoritmu pro budování sémantických reprezentací (českých) vět (Hadacz, 1998, Horák, 1998). V této souvislosti můžeme již nyní počítat s českou elektronickou lexikální databází typu WordNet (Pala, Ševeček, 1999), jež je budována na synonymických řadách a systematicky zachycuje významové vztahy mezi lexikálními jednotkami, konkrétně vztahy synonymie, antonymie, hyponymie, hyperonymie, meronymie, holonymie a další.

V této souvislosti bych rád vyjádřil dík K. Osolsobě, S. Francovi a P. Ševečkovi za obětavou spolupráci, která nakonec vedla do značné míry k úplnému zpracování velkého množství empirických dat. Jde o nesčetné a nepočítané hodiny strávené před obrazovkou, bez nichž by nebylo možno uvedených výsledků dosáhnout. Za práci na budování korpusů instalovaných nyní na Fakultě informatiky MU je potřeba poděkovat P. Rychlému, P. Smržovi, M. Veberovi, A. Horákovi a E. Žáčkové a R. Sedláčkovi z Laboratoře zpracování přirozeného jazyka na FI MU. Za četné připomínky k práci vděčím též prof. dr. P. Maternovi, chyby a nepřesné formulace jsou moje.

Děkuji také dřívějším pracovníkům Ústavu výpočetní techniky Masarykovy univerzity dr. M. Benešovskému, CSc., dr. M. Šmídkovi, CSc. a dr. J. Gerbrichovi za pomoc při zvládnání systému WANDER (Benešovský, Šmídek, 1984) a operačního systému počítače PDP 11, dále pak doc. L. Matyskovi a D. Tomanovi za příspěvní při práci s PROLOGEM a v neposlední řadě také doc. dr. V. Račanskému, řediteli ÚVT MU, za podporu v oblasti technického vybavení i oblastech jiných.

V neposlední řadě bych rád konstatoval, že za řadu východisek a konkrétních podnětů vděčím prof. dr. P. Sgallovi, DrSc. jako svému původnímu školiteli¹. Za podstatná pokládám společná metodologická východiska a zejména pak potřebu nespokojovat se s obraznými, ne zcela určitými, a tedy ne plně kontrolovatelnými formulacemi, pracovat s pojmy definovanými na základě operativních (testovatelných) kritérií a uváděnými do jasných, explicitně formulovaných vzájemných vztahů a konečně nezůstávat u popisu jednotlivých skupin jevů, ale snažit se o zobecnění (Sgall et al, 1985).

Vývoj v oblasti počítačového zpracování přirozeného jazyka se v poslední době zrychluje: při vzniku tohoto textu v r.1993 jsme ještě prakticky neuvažovali o možnosti bezprostředního propojení počítačového zpracování českých textů s podobným počítačovým zpracováním mluveného jazyka, tj. se syntézou a rozpoznáváním mluvené češtiny. Díky příznivému vývoji na Fakultě informatiky, na které začal od r.1996 pracovat doc. Ivan Kopeček orientující se na syntézu a rozpoznávání mluvené češtiny, lze nyní navázat na sebe oba dříve samostatné směry výzkumu a prezentovat je již jako zřetelně integrující se celek. I když kapitoly ... jsou převážně věnovány ... a kapitoly ..., může čtenář v kap. ... najít popis vzájemných vazeb a souvislostí.

0.2 Úvod

Předmětem naší pozornosti je počítačové zpracování přirozeného jazyka (dále PJ). Uvedme několik dobrých důvodů, pro které si PJ zaslouží pozornost:

¹V této souvislosti je třeba uvést, že když jsem v r. 1971-72 dokončoval svou kandidátskou práci, byl mým řádným školitelem prof. dr. P. Sgall. V rámci právě začínající normalizace mi tehdy byl jako školitel odňat a místo něho mi byl *přidělen* doc. dr. J. Popela – i když nemám k dispozici detailní podklady, není obtížné dovést, že se tak nepochybně stalo z iniciativy tehdejšího kompetentního proděkana (děkana) pro vědu na FF UK a možná i její vědecké rady

- jazykové chování představuje jeden z fundamentálních aspektů lidského chování,
- PJ je podstatnou složkou našeho života jako nástroj komunikace,
- jazykové texty slouží jako nosiče pro předávání znalostí z generace na generaci.

Cílem našeho úsilí v této souvislosti je popisovat strukturu přirozeného jazyka tak, abychom na tomto popisu mohli budovat formální (počítačové) modely jazyka, které by vedly k počítačovým programům schopným řešit jednotlivé úlohy zahrnující porozumění přirozenému jazyku. Na konci naší snahy jsou tedy realistické modely takových činností, jako jsou *psaní, čtení, mluvení, poslouchání a vedení dialogu* a další.

Přirozený jazyk se studuje a zkoumá v řadě disciplin, mezi něž patří:

- **lingvistika** – má své vlastní metody a člení se dále na tradiční, klasickou a na metodologicky pokročilejší: **strukturní** či **formální (algebraickou, generativní)** opírající se postupy z oblasti teorie formálních gramatik a jazyků (Chomsky, 1956). Zkoumá vlastní strukturu jazyka, např. prvky, z nichž se skládají slova, dále, jak se slova kombinují do vět, proč některé věty mají určitý význam a jiné nikoli,
- **psychologie, resp. psycholingvistika** – studuje procesy jazykové produkce a porozumění experimentálními technikami, jak lidé rozpoznávají jednotlivé větné konstrukce a jak reagují na významy vět,
- **filosofie a logika** – zkoumá, jak slova mohou něco označovat a jak pomocí jazykových výrazů lze identifikovat objekty v universu promluvy. Zajímá se též o to, co jsou **víry, přesvědčení a komunikační intence** a jak se tyto kognitivní schopnosti vztahují k jazyku,
- **počítačová lingvistika** – klade si za cíl budovat počítačovou teorii jazyka, na rozdíl od klasické lingvistiky se opírá o pojmy **algoritmus, datová struktura** a další – vycházející z počítačové vědy (Computer Science). V počítačové lingvistice se systematicky usiluje o využití poznatků, získaných v jiných oblastech výzkumu, mj. v oblasti AI.
- uvedené samostatné disciplíny lze také zkombinovat do jednoho většího celku a mluvit pak o **kognitivní vědě**. Na některých výzkumných pracovištích (nejčastěji v USA) se můžeme setkat s tímto přístupem.

Je tu přinejmenším dvojí motivace budovat počítačové modely jazyka:

- výzkumná, vědecká, úsilí o lepší pochopení toho, jak funguje přirozený jazyk a jazyková komunikace. Klasické přístupy na to již nestačí, protože ve své tradiční podobě pracují jen s omezenými daty, která lze ještě zpracovat ručně. Proto se nyní pracuje s textovými korpusy obsahujícími stovky miliónů jednotek (obvykle slov). Vznikají programy, které mohou fungovat i jako modely jazykového chování.
- technologická, praktická – počítačové techniky zpracování přirozeného jazyka mohou na druhé straně přinést další revoluci v použití počítačů. V tomto ohledu vzniká nová disciplína – jazykové inženýrství (language engineering), která představuje kombinaci lingvistiky a počítačové vědy a zaměřuje se hlavně na tvorbu programového vybavení pro zpracování PJ.
- potřeba dvoucestné komunikace mezi člověkem a počítačem. Dosavadní komunikační schéma mezi člověkem a strojem je jednocestné a nepřipouští zatím komunikaci lidského typu. Komunikačně bohatší rozhraní v PJ umožní přístup ke složitým počítačovým systémům i neprogramátorům. Systémy s PJ rozhraním by měly být pružnější a inteligentnější než ty dosavadní. Nemusí to nutně být přesné modely lidského uživatele jazyka, hlavním požadavkem ovšem je, aby rozumně fungovaly i pro počítačové nespecialisty. Úspěch v tomto bodě bude mít i rozsáhlé komerční důsledky.

V tomto textu se budeme pohybovat na půli cesty mezi oběma uvedenými možnostmi. Vycházíme přitom z toho, že PJ je natolik složitý, že *ad hoc* přístupy neopírající se o dobře specifikované teorie nemají naději na dlouhodobý a systematický úspěch. Často se však nevyhneme kompromisním řešením, protože naše skutečné znalosti o PJ nejsou vždy na takové úrovni, aby už teď dovolovaly spolehlivě budovat kognitivně přesné a adekvátní modely PJ.

Představu o dané problematice si lze poměrně dobře udělat, když se podíváme na jednotlivé aplikace v oblasti PJ, které se postupně objevují na softwarovém trhu. Celkem zřetelně se vydělují dvě skupiny:

1. programy pro zpracování textů v PJ – sem patří
 - jazyková podpora na úrovni textových procesorů, tj. korektory překlepů (spell checkers), gramatické korektory (grammar checkers), dělicí programy,
 - vyhledávací (fulltextové) programy založené na lemmatizaci (tj. morfologické analýze),
 - programy pro strojový překlad z jednoho jazyka do druhého, obvykle jen pro určité typy textů a experimentální povahy, kvalita překladu nebývá vysoká,
 - prohlížečské programy (browsers) využívající jednoduché morfologické analýzy a klíčových slov, prohlížení e-mailu, dokumentů na WWW.
2. dialogově orientované aplikace, např. dotazovací systémy pro přístup k datovým bázím, automatizované systémy pro komunikaci (i hlasovou, telefonem) s klienty

v bankách nebo knihovnách,
– informační systémy na nádražích a letištích,
– hlasové ovládání počítačů – operační systémy typu Merlin apod., systémy převádějící text na mluvenou řeč (Text-to-Speech Systems, TTS), u nás např. Demosthenes (Kopeček, 1999) a též AUDIS (Kopeček, 1998), dále sem patří systémy pro rozpoznávání mluvené řeči (Automatic Speech Recognition Systems, ASRS) s aplikacemi v podobě diktovacích systémů typu Via Voice (IBM) či Dragon (firma Lernout & Hauspie),
– expertní systémy různého typu, např. diagnostické systémy pro lékaře (MYCIN), automechaniky aj., databázové systémy s rozhraním umožňujícím klást dotazy v PJ,

3. atraktivní oblastí pro textově orientované systémy je **porozumění příběhům** (story understanding). Do tohoto okruhu patří systémy, které dovedou porozumět novinovým článkům a vytvářet z nich souhrny a abstrakty. V USA se každoročně koná testování těchto systémů ve formě soutěže (www.).

Poznámka

Je důležité rozlišit problematiku strojového rozpoznávání řeči (speech recognition) a porozumění PJ. Systém pro rozpoznávání řeči nemusí ještě zahrnovat skutečné porozumění přirozenému jazyku. Např. hlasově ovládané počítače, které se nyní objevují na trhu, nezahrnují porozumění PJ v obecném (lidském) smyslu. Rozpoznávaná slova fungují jen jako příkazy (signály) pro provedení příslušné operace, ale nejde o porozumění ve smyslu typické dvoucestné komunikace mezi lidmi. To dovedou do jisté míry systémy pro porozumění PJ, které by pak mohly mít jako vstup právě výstup z rozpoznávače řeči.

0.2.1 Roviny analýzy jazyka

Systémy pro zpracování PJ se neobejdou bez potřebných znalostí o vlastní struktuře jazyka, musí v nich být zabudovány znalosti o tom:

- – co jsou slova,
- – jak se slova kombinují do vět,
- – co slova označují, jaké jsou jejich významy,
- – jak se význam věty skládá z významů slov.

To však ještě nestačí – inteligentní jazykové chování uživatele jazyka – člověka (dále UJ) se opírá o obecnou (encyklopedickou) znalost světa a jeho inferenční schopnosti a

také o znalost komunikační situace a komunikačního kontextu a pravidel, podle nichž se komunikační procesy řídí.

I když to, co jsme právě uvedli, vypadá na první pohled celkem jednoduše a samozřejmě, skutečnost je podstatně komplikovanější. Znalosti relevantní pro počítačové zpracování přirozeného jazyka (dále ZPJ) mají komplikovanou hierarchickou povahu, proto je obvyklé mluvit v této souvislosti o jednotlivých rovinách popisu, tj. o rovině:

1. **fonetické a fonologické** – postihuje vztahy mezi zvuky a dalšími jednotkami (např. slabikami), z nichž se slova tvoří. Rozlišují se tu **fonémy**, což jsou nejmenší jednotky jazyka schopné rozlišit význam (např. *m* a *t* ve slovech *máme* a *máte* nebo *m* a *n* v *tomu* a *tonu*). Tyto a další znalosti jsou podstatné pro systémy založené na rozpoznávání mluvené řeči,
2. **morfologické** – týká se toho, jak se slova skládají ze základnějších jednotek nazývaných **morfémy**. Jsou to nejmenší jednotky jazyka, které mohou nést význam. To lze demonstrovat na příkladech segmentace výrazů jako *nej-ne-u-věř-i-t-eln-ějš-ího*, *uč-e-n-í*, v nichž rozlišujeme **kořeny**, **kmeny**, **kmenotvorné přípony**, **prefixy**, **suffixy**, **koncovky**. Ve flektivních jazycích, jako je čeština, jsou morfologické vztahy bohatě rozvinuty – vyznačují se komplikovanou deklinací (skloňováním) a konjugací (časováním). Ohýbání slov je potřeba algoritmicky popsat a na tomto základě vytvořit vhodné analyzátory a generátory tvarů.
3. **syntaktické** – vysvětluje, jak lze spojovat slova tak, aby z nich vznikaly **gramaticky správné věty**, z jakých prvků, **složek** se skládají věty a jaké mezi nimi existují vztahy a jak lze tyto vztahy formálně reprezentovat. Na základě těchto znalostí je pak možno budovat syntaktické analyzátory a generátory, což jsou v konečné fázi počítačové programy, které na vstupu přijímají věty přirozeného jazyka a na výstupu poskytují jejich reprezentace nejčastěji v podobě stromových struktur (grafů-stromů).
4. **sémantické** – popisuje, co jazykové výrazy (slova a jejich spojení, kolokace) znamenají a jak se jejich významy kombinují tak, aby tvořily **smysluplné** (sémanticky dobře utvořené) věty. V tomto bodě uvažujeme významy vět **nezávisle na kontextu**. I zde celkově usilujeme o vytvoření sémantických analyzátorů, tj. v konečném úhrnu programů, které vstupním větám přirozeného jazyka budou přiřazovat jejich sémantické reprezentace mající podobu symbolického formálního zápisu, např. to mohou být formule v predikátovém kalkulu 1.řádu nebo lépe formule lambda kalkulu, jestliže se rozhodneme použít transparentní intenzionální logiky (TIL, Tichý, 1989, Materna, 1999).
5. **pragmatické** – tj., jak se vět užívá v různých komunikačních situacích (sdělení, rozkaz, otázka, přání, slib, prohlášení – deklarace nezávislosti) a jak užití vět ovlivňuje interpretaci jejich významu.

6. kontextové, promluvové – zachycují, jak bezprostředně předcházející věty ovlivňují sémantickou interpretaci vět následujících, např. v promluvě *Naši si koupili dům a auto. To vedlo k velkým nepříjemnostem.*
7. patří sem i znalosti o světě, které zahrnují obecné encyklopedické znalosti, jimiž uživatel jazyka musí disponovat, aby byl schopen vést normální komunikaci. Ve skutečnosti jde o složitý komplex znalostí, k nimž se řadí též znalosti o komunikačních záměrech, plánech a vírách ostatních uživatelů jazyka a v neposlední řadě i znalosti a soubory inferenčních pravidel označované jako zásady zdravého rozumu (common sense).
8. Vyčlenit je potřeba i jazykové metaznalosti, které propojují znalosti o světě se znalostí daného přirozeného jazyka.

Uvedený výčet se jeví jako základní rámec znalostí potřebných pro zpracování PJ: algoritmy pro ZPJ, které si činí nárok na jistou míru obecnosti, musí zahrnovat kombinace znalostí současně z několika rovin, takže míra jejich složitosti je pak vysoká. Pro další výklad se přidržíme naznačeného rámce.

0.2.2 Reprezentace a porozumění

Klíčová složka porozumění spočívá podle našeho názoru ve vybudování reprezentace významu vět a textů. K tomu je však třeba definovat, co je to reprezentace významu.

První – přirozenou – možností, která se nabízí, je: věty samy by mohly sloužit jako reprezentace svého významu. Proti tomu stojí argument, že slova, jazykové výrazy jsou víceznačné, mají více významů (smyslů), viz např. výrazy jako *kopu, je, červená* a také výrazy jako *hlava, strana, stát, dostat, mít* aj. Tato víceznačnost velmi komplikuje možnost vyvozovat formálně vhodné inference, bez nichž se model porozumění neobejde.

Pro UJ – lidi nepředstavuje zjednoznačňování, desambiguace jazykových výrazů obtížný problém, děláme ji automaticky, podvědomě. Lidští UJ obvykle neuvažují zvlášť každý jednotlivý význam, když rozumí větám, když je chápou. Algoritmický popis porozumění, program na něm založený to však dělat musí, musí být explicitní.

Tato úvaha vede k závěru, že pro reprezentaci významu potřebujeme jiné prostředky než přirozený jazyk. Co se tedy nabízí? Dosavadní výzkumy se shodují v tom, že vhodným nástrojem pro reprezentaci významu má být nějaký formální (matematický, logický) jazyk, tj. symbolický jazyk, jehož základními prvky jsou atomické symboly a na jehož výrazy lze aplikovat princip kompozicionality, který říká, že význam věty, jazykového výrazu lze přirozeným způsobem složit z jeho složek.

Existuje obecná shoda v tom, že vhodný jazyk pro sémantickou reprezentaci vět a výrazů přirozeného jazyka by měl mít následující vlastnosti:

1. reprezentace významu musí být **přesná a jednoznačná**, tj. pro každý samostatný význam musí také existovat samostatná reprezentace, tedy **samostatná formule**, ev. term či podformule.
2. reprezentace by měla zachycovat intuitivní strukturu vět (výrazů) přirozeného jazyka. Věty podobné svou strukturou by měly být reprezentovány strukturně podobnými reprezentacemi.
3. významy dvou vět, které jsou **vzájemnými parafrázemi**, tj. mezi nimiž existuje vztah synonymie (antonymie), by také měly být k sobě vztaženy prostřednictvím svých reprezentací.
4. reprezentace významu by měla být pokud možno **nezávislá** na daném přirozeném jazyce.

Na tomto místě je třeba zdůraznit, že pro jednotlivé výše uvedené úrovně je díky jejich odlišnosti počítat s **různými** reprezentacemi, jinými slovy, každá rovina má svou vlastní reprezentaci, tj. svou vlastní **formální notaci** pro zachycení příslušné reprezentace. Rozumný NLP systém musí být schopen tyto reprezentace propojit a navázat na sebe v jednom složitém formálním systému.

V dalším se pokusíme naznačit, jak formálními prostředky reprezentovat:

- **morfologické struktury**: jsou konstituovány slovy a jejich součástmi – **morfémy**, nejmenšími jednotkami jazyka, které jsou schopny nést význam. U systémů pro porozumění potřebujeme rozpoznat **morfémovou strukturu** slov(a) nebo, což je prakticky totéž, provádět morfologickou analýzu slov ve vstupním textu, ev. jejich syntézu, tj. generovat všechny přípustné slovní tvary. Lze to dobře ilustrovat na českém tvaru jako *nej-ne-po-chop-i-t-eln-ějš-ího*: rozpoznání (segmentace) jeho morfémové struktury spočívá v identifikování **kořene**, který obvykle definujeme jako morfém nesoucí lexikální význam, a dalších morfémů – **prefixů** a **sufixů**, které obvykle nesou významy gramatické – **tvarotvorné**, **slovotvorné** nebo některé modifikující významy lexikální, např. *-eln-* – ”ten, který je možno...”. V jazyce, jako je čeština, je kombinatorika morfémů do značné míry **pravidelná**, a proto i systematicky popsatelná souborem formálních pravidel, která z gramatik známe jako **vzory**, a to vzory **deklinanční** postihující ohýbání substantiv, **konjugační** popisující ohýbání sloves a ostatní – zachycující třídy neohebných slov – i pro ně se vyplatí zavést jejich vlastní vzory. Hledáme-li formální prostředky, které umožňují vhodně (i z hlediska implementačního) reprezentovat morfémové struktury českých slov, ukazuje se, že k tomuto účelu mohou dobře složit některé typy **konečných automatů** a **trie** struktury – tohoto přístupu je použito v morfologickém analyzátoru a lemmatizátoru pro češtinu LEMMA (Ševeček, 1995) a zejména v morfologickém programu AJKA podrobně popsaném v práci (Sedláček, 1999). Detailněji se této problematice budeme věnovat níže.

- **syntaktické struktury (vět):** postihují vztahy mezi prvky (slovy), z nichž se věty či rozsáhlejší jazykové výrazy skládají. Jinak řečeno, pomocí syntaktických struktur reprezentujeme stavbu vět a jazykových výrazů, zachycujeme jimi, jak se jednoduché (atomické) větné složky (obvykle slova) seskupují do větších celků, jak jedny větné složky **modifikují** druhé, **vyznačují**, které výrazy jsou ve větě nejzávažnější – gramaticky i významově. Mějme např. věty

- (1) *Honza prodal ten počítač Petrovi.*
- (2) *Počítač byl prodán Petrovi (Honzou).*
- (3) *Počítač se prodal (někdo někomu).*

Tyto věty sdílejí určité strukturní i významové (sémantické) vlastnosti, které by měly být v reprezentaci zachyceny. V obou větách jde sémanticky o činnost *prodávání*, přesto se však v jistém podstatném ohledu od sebe liší.

Když se podíváme na věty jako

- (3) *Honza dal knihu.*
- (4) *Eva jsou v kuchyni.,*

je zřejmé, že jsou určitým způsobem **neúplné**, **deviantní**. Můžeme o nich říci, že nejsou **gramaticky správné**. I toto je potřeba v reprezentacích syntaktických struktur vhodným způsobem zachytit.

Pak jsou tu případy jako

- (5) *Hutě železa vyrábějí málo.*
- či

- (6) *Kritika poslanců vedla k rozpadu koalice.*

Je vidět, že každá z těchto uvedených vět dává dvě různá čtení, která bychom chtěli vhodným způsobem reprezentovat, tj. zachytit je v našich zamýšlených syntaktických reprezentacích. Nejčastěji se k tomuto účelu užívá **stromových struktur**, resp. **grafů-stromů** (frázových ukazatelů, strukturních popisů opírajících se o formalismus nekontextových gramatik), které reprezentují větné struktury v termínech jejich složek. Pro věty (1) a (2) můžeme mít reprezentace jako (1a) a (2a). Existuje také možnost pracovat se **závislostními grafy** – té zde nevyužíváme (viz. např. Hajičová, PDTB Grafy mohou vypadat následovně:

- (1a)
- (2a)

- **významy slov a významy vět** – reprezentace významu: syntaktické reprezentace neodrážejí přímo význam vět, zachycují ale vztahy, které jsou klíčové pro rozpoznání jejich plného významu. V příkladech jako (5) a (6) potřebujeme rozlišit různá čtení nezávisle na kontextu a potřebujeme to udělat vhodnými formálními prostředky tak, aby jednotlivá čtení byla explicitně rozlišitelná.

To lze udělat třeba tak, že najdeme způsob, jak reprezentovat sémantické vztahy mezi slovesem a jeho doplněními nebo jinými slovy, významové vztahy mezi predikátem a jeho argumenty (např. np, pp, adg, s). Věty (1) a (2) pak můžeme zkusit reprezentovat např. takto:

(1b) $\text{prod}(\text{ag}, \text{obj}, \text{adr})$,

kde ag interpretujeme jako **agens**, činitel (ten, kdo něco dělá), obj jako **objekt**, který se prodává (co je činností zasaženo, co z ní vzniká), a adr jako **adresát**, ten, komu je určen **objekt**

nebo (1c) $\text{prod}(\text{kdo}, \text{co}, \text{komu})$,

kde použité zájmenné výrazy lze interpretovat prakticky stejně jako výše. Tento způsob zachycuje, o co nám jde, totiž že věty (1) a (2) se neliší významově, ale jen **povrchově**, jiným uspořádáním syntaktických vztahů, jejich jinou perspektivou. Budeme-li chtít věty (1) a (2) reprezentovat jako znalost vyjadřující, že nějaký konkrétní počítač změnil majitele, můžeme odpovídající fakt reprezentovat ještě jinak:

(1d) $\text{prod}(\text{h3}, \text{poč13}, \text{p5})$,

kde prod lze interpretovat jako logický predikát označující vztah prodávání a h3 , poč13 , p5 jeho odpovídající argumenty, v tomto případě individuální konstanty referující k příslušným objektům v universu promluvy. Chápeme-li (1d) jako logický predikát, pak to znamená, že jsme se rozhodli význam vět (1) a (2), ale i dalších reprezentovat pomocí aparátu PK1, který má některé výhody a řadu nevýhod, o nichž se zmíníme později. Mezi jeho výhody patří:

- je dobře formálně propracován a definován,
 - existuje řada zkušeností s jeho použitím, viz např. SHRDLU (Winograd, 1974), LUNAR (Woods, 1976), KRL (), CYCORP (1995),
 - existuje pro něj počítačová implementace ve formě programovacího jazyka PROLOG (vyvinutého mimochodem pro potřeby NLP, Colmerauer 1979.).
- plnou reprezentaci významu vět je možno spolehlivě získat jen s přihlédnutím ke znalostem o světě, jež jsou dnes v systémech pro porozumění PJ zachycovány pomocí speciální **reprezentace znalostí**. Jde o notační systémy podobné reprezentaci významu uvedené výše, tj. systémy založené na PK1 nebo na transparentní intenzionální logice (systému TIL, Tichý 1989, Materna 2000, Hadacz 2000, Hadacz, Horák, 2000). V dosavadních výzkumech lze pozorovat poměrně striktní oddělování reprezentace významu od reprezentace znalostí, které plyne z potřeby provádět nad reprezentací znalosti potřebné inference umožňující odvozovat z jedněch fakt jiná. Je však vidět, že reprezentace znalostí v dosavadních podobách postrádá propracovanou návaznost na to, čemu se obvykle říká **encyklopedické znalosti** a také na **é metaznalosti**, jež zahrnují speciální znalosti o jazyce, jednotlivých jazykových výrazech a jejich kolokabilitě. Zejména dosavadní elektronické slovníky jsou budovány příliš úzce a nebere se v nich zřetel na těsné souvislosti mezi jazykovými a encyklopedickými znalostmi.

Typická struktura NLP systému – obr. a komentář. Vstupní věty jsou nejprve podrobeny lexikální analýze využívající **slovníku**, který obsahuje znalosti o významech slov, pak morfologické a syntaktické analýze opírající se o množinu pravidel definujících

přípustné syntaktické struktury – tedy o **gramatiku**: to vše v modulu, který se obvykle nazývá **parser (analyzátor)**. Získané syntaktické reprezentace jsou pak sémanticky interpretovány a výsledkem jsou sémantické reprezentace – zde, jak patrně, v PK1. V poslední době se však místo sekvenční strategie analýzy preferují postupy **paralelní (rule-to-rule)**, kdy každému syntaktickému pravidlu v gramatice odpovídá příslušné pravidlo sémantické, které se provádí pokud možno souběžně. Tím se značně redukuje počet možných interpretací a také to pravděpodobně lépe odpovídá povaze lidského porozumění větám PJ.

Máme-li věty:

(7) *Návštěvy příbuzných jsou únavné.*

a

(8) *Návštěvy muzeí jsou únavné.,*

vidíme, že jejich odpovídající syntaktické struktury jsou syntakticky víceznačné, obě varianty jsou platné, ovšem k rozhodnutí, kterou z nich vybrat, je nutná znalost kontextu (kdo koho navštěvuje, a také kdo koho může navštěvovat, což je de facto znalost o světě). Právě proto je u věty (8) možná jen jedna sémantická interpretace (muzea mohou sotva někoho navštěvovat). Při použití sekvenční strategie se u věty (8) nevyhneme pokusu o dvojí sémantickou interpretaci, zatímco při souběžné aplikaci syntaktického a sémantického pravidla a přihlédnutí k encyklopedickým znalostem by už k vybudování druhé syntaktické struktury nemělo dojít, zjištěná možnost sémantické anomálie by měla další pokusy eliminovat. V tomto příkladě se vyhneme jedné zcela chybné sémantické interpretaci, ovšem u reálných aplikací se setkáváme s větami připouštějícími řádově více než několik desítek syntaktických struktur, z nichž většina pak vede k sémanticky nekorektním interpretacím.

Povšimněme si ve schématu modulu označenému jako **kontextová interpretace (analýza promluvy)**. Je to proces, který zahrnuje přinejmenším následující procedury:

- identifikaci objektů označovaných jmennými skupinami (*ten nový počítač*), zájmeny (*ty, on, tu, teď*) a na ni navazující referenčních a koreferenčních vztahů,
- temporální zařazení informace nesené danou větou ve vztahu k okamžiku promluvy,
- identifikaci postoje mluvčího, např. zda ve větě *Je tady chladno.* jde o konstatování faktu nebo rozkaz (žádost),
- inference potřebné k náležité interpretaci věty v rámci dané aplikační oblasti – na základě znalosti předchozího kontextu (předcházejících vět) a dané aplikační oblasti (třeba počítače a politika), viz věty jako *Programátor zavedl nový operační systém.* a *Vláda sociálních demokratů zavedla nové daně.*

0.3 Data pro zpracování přirozeného jazyka – korpusy

Jazyková data mají empirickou povahu, a proto je zjevné, že úspěšnost popisu přirozeného jazyka je do značné míry závislá na tom, jaký máme přístup k datům a v jaké podobě jsou nám jazyková data k dispozici. Protože většina jazykových dat má podobu textů (psaných nebo písemně zachycených (transkribovaných) mluvených), je možnost mít je pohromadě v elektronické podobě klíčová pro další rozvoj lingvistiky a zpracování přirozeného jazyka vůbec. Korpusová lingvistika v současnosti představuje novou větev lingvistiky, v níž se pracuje s korpusy uloženými v počítačích. To přirozeně znamená, že se v mnoha aspektech překrývá s počítačovou lingvistikou, z níž čerpá řadu postupů a technik. Výsledky získané těmito postupy brzy výrazně ovlivní nejen samu lingvistiku, v níž si jistě vynutí vznik nových, úplnějších a empiricky adekvátnějších gramatik (v knižní podobě), ale i počítačové zpracování přirozeného jazyka jako celek – už dnes se na základě korpusových dat budují nové a přesnější elektronické slovníky a robustní počítačové gramatiky. Korpusy jsou dnes v jazykovém inženýrství východiskem pro realistický základní výzkum ve formě relativně blízké přírodním vědám.

Není těžké vidět, že symbióza korpusové lingvistiky s počítačovou má i jasné metodologické důsledky: lingvista dnes může dělat věci, které byly dříve nepředstavitelné ať už pro svou časovou náročnost a pracnost (viz např. jednoduchý úkol seřadit manuálně třeba 250 tisíc slovníkových hesel) nebo skutečnou složitost (např. nalezení všech výskytů předložky *na* spolu se substantivem v akuzativu v textech o rozsahu 100 mil. slovních tvarů – spojení jako *na stůl*, *na týden*). Jedním z důsledků je i to, že lze systematictěji využívat statistických a pravděpodobnostních přístupů, které by se bez počítačů na velké soubory nedaly aplikovat.

Korpusy nejsou určeny jen pro lingvisty – přirozený jazyk je prostředkem komunikace pro všechny: proto jejich budování není jen záležitostí lingvistů a jazykových inženýrů. V našich podmínkách lze konstatovat, že pochopení tohoto prostého faktu se pozitivně projevilo tím, že díky příspěví GA ČR se v rámci komplexního grantového projektu K214 (*Čeština ve věku počítačů*) začal budovat Český národní korpus čítající aktuálně cca 200 mil. českých slovních tvarů. Vedle toho byl v rámci projektu VS97028 (Program 250 – podpora výzkumu na VŠ, MŠMT ČR) na FI MU vybudován další samostatný obecný korpus ESO, který v současnosti čítá cca 160 mil. českých slovních tvarů. V tomto ohledu jde o jasně interdisciplinární záležitost, neboť korpusová data jsou použitelná pro odborníky v řadě disciplin:

- sociology a sociolingvisty,
- psychology,
- odborníky v oblasti masové komunikace a médií (reklama),

- lexikografy a lingvisty, překladatele (strojový překlad),
- výzkumné pracovníky v oblasti umělé inteligence (porozumění přirozenému jazyku, reprezentace znalostí, robotika aj.),
- tvůrce učebnic a tzv. referenčních příruček (gramatiky, slovníky)

V současnosti se korpusem rozumí rozsáhlý vnitřně strukturovaný a ucelený soubor textů daného jazyka elektronicky uložený a zpracováváný. Dnes vytvářené korpusy jsou organizovány se zřetelem ke zvolenému cíli (pro potřeby lexikografů, sociologů, komunikačních odborníků) a vycházejí z následujících teoretických předpokladů:

1. jazyková data jsou v korpusu uložena ve své **přirozené** textové podobě, proto je lze všestranně a opakovaně zkoumat a vyvozovat z nich příslušné teoretické generalizace,
2. **velký rozsah dat** v korpusu minimalizuje nebezpečí, že by mohlo dojít – třeba i náhodou – k převaze okrajových jevů nad základními a typickými,
3. **velký rozsah dat** v korpusu je podmínkou dostatečné **reprezentativnosti**, což např. při budování slovníků vůbec nemusí být jednoduchá záležitost:

Lze to ukázat na vztazích mezi pojmy: **token** (výskyt), **typ** a **lemma**. Token chápeme jako výskyt slovního tvaru v korpusu, typ – slovní tvar jako takový a lemma je základní tvar pro nějakou skupinu tvarů (např. nominativ u substantiv nebo infinitiv u sloves). Uvedme proporce těchto entit v *BNC* pro psaný jazyk:

tokens: 90 miliónů (v *BNC* je 10 mil. tvarů z mluveného jazyka)

typy: 524 060

– z toho typů s četností 1 je: 258 575

– 2% typů pokrývá 90% výskytů (tokens)

lemmata: proporce typ : lemma, např. v SOD (Students Oxford Dictionary), činí pro angličtinu 2,5 : 1. Z uvedených údajů lze odvodit, že např. pro slovník, který by měl mít rozsah cca 250 tis. heslových slov, poskytuje *BNC* se svými 100 mil. slovními tvarů reprezentativní materiál jen pro cca 100 tis. heslových slov (de facto lemmat).

0.3.1 Jak se budují korpusy?

Zdrojem korpusových dat je jak jazyk psaný, tak i mluvený, u dosavadních korpusů to bývá zhusta v poměru 9:1, protože záznam mluveného jazyka (magnetofonová nahrávka) a jeho převod (manuální přepis) do počítačově čitelné podoby je zatím velmi nákladný (až 15krát dražší než u psaných textů). Situace se může zlepšit až s komerčními aplikacemi pro zpracování mluvené řeči.

Z psaných textů se data získávají prakticky třemi způsoby:

- konverzí ze sázecích disket a pásek, které lze získat od většiny nakladatelství vydávajících noviny, časopisy a knihy,
- užitím technik OCR, jejíž úspěšnost je do značné míry závislá na kvalitě použitého scanneru a programového vybavení a na typografické složitosti textu – typech a velikostech písem,
- klasickým manuálním opisováním textů do počítače.

Ve všech případech je nutná kontrola, opravy chyb, ev. konverze mezi použitými a typicky odlišnými kódy. Výsledek se zpravidla ukládá do mezinárodního ASCII formátu – ovšem v případě češtiny je třeba mít k dispozici vhodné konverzní programy, protože čeština je kódována řadou způsobů: (v kódech MJK, PCL2, IL2 a 1250 ve Windows).

V neposlední řadě se při tvorbě korpusů tvůrci musí vyrovnávat i s právními aspekty objevujícími se při získávání dat. Týká se to copyrightu a autorských práv a jejich uvolnění ze strany autora či vydavatele. Jednodušší bývá situace v případě nekomerčního využití, jinak je potřeba uzavírat vhodné typy smluv přesně stanovujících podmínky šíření korpusových dat a produktů, které na jejich základě vznikly. U mluvených záznamů je často potřeba zajistit zachování anonymity mluvčích.

0.3.2 Typy korpusů a standardizace

Textové soubory volně uložené v počítači ještě netvoří korpus. Obvykle se setkáváme s následujícími typy uložení jazykových dat:

- elektronické archivy – volné kolekce celkově různorodých textů. Klasickým příkladem je *Oxford Text Archive – OTA*, který představuje rozsáhlou sbírku různých, většinou literárních textů, v různých formátech a různých jazycích: v OTA najdeme asi tisícovku literárních textů v 25 jazycích a různých formátech,
- vlastní korpusy tvořící relativně úplné celky, i tak ovšem značně různorodé a lišící se v řadě parametrů,
- podle jazyků – dnes už jen málo jazyků v Evropě nemá svůj korpus, v r. 1990 existovaly korpusy pro:
 - angličtinu:220 000 000 slovních tvarů (a 20 korpusů)
 - francouzštinu: 190 000 000 slovních tvarů
 - němčinu:27 500 000 slovních tvarů
 - holandštinu: 60 000 000 slovních tvarů
 - italštinu:30 000 000 slovních tvarů
 - srbochorvatštinu: 12 000 000 slovních tvarů
 - korpusy dvoujazyčné, paralelní: anglicko-francouzské, -italské, -dánské

– korpusy obecné a specifické, velké obecné korpusy obsahují subkorpusy jazyka psaného, mluveného, nářečí, synchronní – diachronní aj.

S rostoucím počtem korpusů vzniká potřeba jejich standardizace a vícenásobného a sdíleného použití (jedna z důležitých podmínek v rámci EU). S tímto cílem vznikla Text Encoding Initiative – TEI sponzorovaná EU a americkou vládou: vydala již doporučení pro společný výměnný formát, zásady kódování, znakové sady a navrhla společný kódovací – značkovací meta-jazyk, jímž je Standard Generalized Markup Language – SGML, určitě známý některým uživatelům TeXu a od r. 1986 uznávaný jako mezinárodní standard (ISO 8879). Značný důraz se klade na polyfunkčnost a polyteoretičnost notace (nezávislost na dílčích teoriích), aby se v budoucnu nemusely dělat nákladné úpravy a změny. Nejnověji se začíná pracovat s jazykem XML (citát), který vychází ze SGML.

0.3.3 Budování korpusu – sběr dat

Na příkladu *BNC* naznačíme jen zhruba základní vlastnosti, které je třeba uvážit při budování korpusu. Nebudeme se pouštět do podrobností, chceme poskytnout jen základní představu. Korpus typu *BNC* může vypadat zhruba takto:

- je to výběrový korpus, tj. skládá se z vzorků ne delších než 40 000 slov, které jsou vybrány v následujících proporcích:

1. přírodní vědy a čistá věda	5%
2. aplikované vědy	5%
3. sociální vědy	15%
4. politická publicistika	15%
5. publicistika obchodní a finanční	10%
6. publicistika umělecká (rock & pop, divadlo,...)	10%
7. publicistika náboženská a filosofická	5%
8. publicistika zábavná (sport, zahrádkáři, ...)	15%

Podíl těchto textů se pohybuje v rozmezí 70-80%, podíl uměleckých textů činí 20-30%.

Další rozlišení se týká toho, zda vzorky pocházejí z knih, deníků, časopisů, dopisů apod.:

1. knihy	55-65%
2. periodika	20-30%
3. brožury, letáčky, příručky, reklamy	5-10%

- 4. dopisy, memoranda, zprávy, eseje5-10%
- 5. mluvené texty7-10%

- je **synchronní**, tedy obsahuje výhradně texty ne starší než např. od r. 1987, a vždy se uvádí datum, kdy byl text publikován poprvé,
- je **obecný** čili není specificky orientován na nějakou konkrétní oblast nebo žánr a zahrnuje vzorky od všech věkových skupin, viz výše,
- je **jednojazyčný** – obsahuje jen vzorky pocházející od anglických (českých,...) mluvčích.
- jsou zavedeny **klasifikační rysy**, které nesledují vyhraněné proporce a jsou orientovány na pozdější využití korpusu (lze podle nich třídit a vyhledávat v celém korpusu):
 1. identifikátor vzorku
 2. rozsah vzorku (počet slov), začátek a konec vzorku
 3. rozsah textu příslušného typu (počet slov)
 4. kompozice textu (hladký, složený, sbírka)
 5. standardní bibliografický odkaz
 6. datum vzniku
 7. předmětná oblast
 8. úroveň složitosti textu
 9. autorství (individuální, společné, institucionální, neznámé)
 10. pohlaví autora
 11. věková skupina autora
 12. etnická skupina autora
 13. autorovo bydliště
 14. věk cílové skupiny (na kterou je text orientován)

0.3.4 Vnitřní struktura korpusu

Vnitřní struktura korpusu

- 1) atributy poziční
- 2) atributy strukturní (hranice vět, odstavců)

slovo	lemma	gr.značky	sém.značky
ženu	hnát/žena	k5/k1gFnSc1	HUM+FEM/POHYB
ovce	ovce	k1gFnPc4	ANIM
na	na	k7c4	DIRECT
pastvu	pastva	k1gFnSc4	LOC

0.3.5 Korpusové nástroje

Problematika korpusových nástrojů je rozsáhlá a představuje pole, na kterém se setkávají požadavky uživatelů (hlavně lingvistů a lexikografů) s přístupy programátorů. Výsledkem je konkrétní programové vybavení umožňující získávat z korpusů „poklady“, které jsou v nich skryty.

Základem jsou obvykle konkordanční programy (např. MicroOCP), které třídí a počítají objekty nalezené v korpusu, což jsou v *syrovém* korpusu slovní tvary, interpunkce, případně další znaky (vyznačující třeba hranice vět, odstavců aj.) – ty jsou typicky součástí SGML. Pokud není do korpusu nějak zavedena další informace, konkordanční program nemůže rozlišit určité víceznačnosti (homonymie), např. v češtině mezi tvary *ženu* (ak.sg. substantiva *žena*) a *ženu* (1.os.sg.prés. slovesa *hnát*), nemluvě již o tom, že tvar *hnát* může být také tvarem substantiva mužského rodu. Proto ke korpusovým nástrojům patří i programy, které představují svého druhu gramatické analyzátoři: orientují se na morfologii, syntax a v poslední době i na sémantiku. V současné terminologii se obvykle mluví o značkování (tagging) a o značkovacích programech (taggers) různé úrovně. Níže uvedené taggery obvykle pracují tak, že se snaží každému slovu v korpusu přiřadit jeho gramatickou značku, tj. jeho slovní druh včetně relevantních gramatických kategorií. Programy uvedené dále buď s těmito analyzátoři spolupracují, nebo je přímo obsahují jako svou součást, nicméně pro přehlednost se o nich dále zmiňujeme zvlášť. **Korpusové manažery** Jako vhodný příklad může posloužit korpusový procesor CQP (Corpus Query Processor), který se vyznačuje následujícími rysy (viz níže):

- vlastní procesor CQP (Christ, Schulze, 1995), implementován v jazyce C, užívá X-Windows, na platformě OS Solaris a Linux,
- uživatelsky přitulnější rozhraní fungující jako nadstavba nad cqp: xkwic a gcqp (Rychlý, Skoupý, 1998),
- zadávání vyhledávacích dotazů funguje na bázi regulárních výrazů,
- výstup: konkordanční seznamy, výskyty slov v kontextech,
- lze vyhledávat kolokace (slovní spojení),
- lze získávat základní frekvenční údaje ke slovům a kolokacím,

- lze počítat další statistické parametry jako MI a T-score,
 - u značkováného korpusu lze vyhledávat podle gramatických kategorií a lemmat a také podle strukturních značek.
1. program: **korpusový procesor cqp** – vytvořen v IMS na universitě ve Stuttgartu, napsán v jazyce C, běží na Sunech (OS Solaris) a pod Linuxem v X-Windows, patří k němu i jeho nadstavba **xkwic**. **cqp** a **xkwic** umožňují v korpusu vyhledávat:
 - výskyty jednotlivých slov spolu s kontexty, v nichž se vyskytují, např. *ovšem* – výsledkem je konkordanční seznam
 - kolokace, např. *ten, který* a také konkordanční seznam
 - základní frekvenční údaje ke slovu
 - dotazy na vyhledání se zadávají pomocí regulárních výrazů, např. požadavek na vyhledání slova *následkem* se zadá: ...
 - podle tzv. **pozičních** a **strukturních** atributů: tj. podle **slov, lemmat a gramatických kategorií**, a pak i podle struktury textu – vět, odstavců apod. – ukázky práce s **cqp** a **xkwic** formou jednoduchých cvičení, vyhledání konkrétních slov a kolokací a využití k dalšímu výzkumu

0.3.6 Značkování (anotování) korpusů

Gramatické značkování (anotování)

Co to je značkování:!!! Věnujme nyní pozornost značkování. Pro příklad vezměme systém, který provádí v korpusu značkování (tagging) slov. Lingvista nejprve navrhne soubor gramatických značek – symbolů reprezentujících slovní druhy, pak souběžně následuje vytvoření slovníku kmenů (slovních základů) a na něj navazující morfologický analyzátor, který na základě segmentace každému výskytu slova v korpusu přiřadí symbol (značku) jeho slovního druhu – což je postup vhodný pro většinu evropských jazyků včetně češtiny.

Předpokládaná úspěšnost takového značkování je do 90 %, chyby, jichž se program dopustil, jsou analyzovány a na základě této analýzy je doplněn slovník kmenů a modifikován analyzátor. Pak lze přikročit k dalším testům a v případě vyšší míry úspěšnosti i k další analýze korpusu. Pro angličtinu se dnes převážně užívá pravděpodobnostního přístupu, pro jazyky typu češtiny se jako vhodnější jeví morfologické analyzátory (viz dále).

Zmínili jsme se už o **gramatickém značkování** (tagging) – přiřazení (symbolů) značek slovních druhů každému výskytu slova v korpusu. Výsledkem je tedy **anotovaný** korpus, tj. ne již čistý (**surový**) korpus, ale jeho verze opatřená gramatickými informacemi jistého druhu.

Takto anotovaný korpus se stává odrazovým můstkem pro další výzkum: pomocí konkordančního programu v něm můžeme vyhledávat gramatické abstrakce, jako

např. výskyty pasíva (seznamy tvarů jako *dělán, prodán, vyroben*), vidu (aspektu) (seznam všech dokonavých sloves s předponou *vy-*), různé posloupnosti slovních druhů aj. Anotovaný korpus poskytuje též výchozí statistická data pro pravděpodobnostní zpracování jazyka. Ke značkováným korpusům patří *Brown Corpus, Lancaster- Oslo-Bergen Corpus (LOB)* a *Spoken English Corpus*, který obsahuje fonetické a fonémické značkování.

0.3.7 Značkování pro češtinu – LEMMA, AJKA

Situace v češtině je jiná než např. v angličtině a podobných jazycích, kde tagger může být jeden program (CLAWS). U nás je potřeba značkování rozložit do dvou fází:

- zpracování morfologickým analyzátozem – morfologická analýza
- desambiguace – manuální, program DESAMB, (ukázat)
 - na bázi partial parsingu – DES
 - statistické techniky (Pary) – techniky strojového učení (Popelinský, Nepil, Žáčková, 2000).

0.3.8 Morfologické (gramatické) značkování

V jazycích, jako je čeština, představuje morfologická analýza samostatný a komplikovaný problém, který se řeší budováním samostatných morfologických analyzátorů (lemmatizátorů) – pro češtinu v současnosti existují tři: LEMMA (Ševeček, Osolsobě, 1995-96) a Hajičův (Hajič, 2000, viz WWW-stránky na MFF UK). Nejnověji se v LZPJ na FI MU pracuje s morfologickým analyzátozem a lemmatizátorem AJKA, jehož autorem je R. Sedláček (Sedláček, DP, 1999).

1. popis AJKY a její činnosti: ukázat interaktivní i dávkové použití Příklad standardního výstupu z programu AJKA (včetně víceznačných tagů):

```
Václav <l>Václav <c>k1gMnSc1
Havel <l>Havel <c>k1gMnSc1
přišel <l>přijít <c>k5eApMnStMmPaP,k5eApInStMmPaP
naopak <l>naopak <c>k6xMeA
s <l>s <c>k7c7
vlastním <l>vlastní <c>k2eAgMnSc67d1,k2eAgXnPc3d1,k2eAgUnSc67d1
      <l>vlastnit <c>k5eAp1nStPmIaI
volebním <l>volební <c>k2eAgMnSc67d1,k2eAgXnPc3d1,k2eAgUnSc67d1
programem <l>program <c>k1gInSc7
,
```

který <l>který <c>k3xQgMnSc15,k3xQgInSc145
nikomu <l>nikdo <c>k3xNnSc3
neubližuje <l>ubližovat <c>k5eNpMnStPmTaI,k5eNp3nStPmIaI

2. morfologická analýza pro češtinu – její principy
3. soubor značek – jeho popis
4. úspěšnost ajky, typy chyb
5. problém víceznačnosti a desambiguace

Cvičení 1): desambiguace v rozsahu 2-3000 slovních tvarů

Cvičení 2): oprava vybraného souboru s chybami v rozsahu...

0.3.9 Syntaktické značkování

Značkování na úrovni vyšší než slovnědruhovému, tj. na rovině syntaktické, lze najít např. v *London-Lund Corpusu* (Svartvik, 1990). Vznikly již syntakticky analyzované subkorpusy známé jako *stromové banky* (*treebanks*), byly však vytvořeny jen z podčástí korpusů. I tak jde o texty v rozsahu několika miliónů slov a o práci, která např. v UCREL zabrala kolem 5 let. Nedávný výzkum na *LOB Corpusu* však vedl k technice zjednodušené syntaktické analýzy známé jako *skeletonová analýza*, kterou lidští operátoři mohou provádět poměrně rychle (Leech and Garside, 1991). Pražský závislostní korpus byl celý vytvořen manuálně.

Stromové banky (*treebanks*)

Jsou to textové soubory tvořené větami, u nichž je vyznačena syntaktická struktura, např. ve tvaru syntaktického (složkového) stromu (ohodnoceného uzávorkováním).

(1) Věděl jsem, že přijde a že mi dá pusu.

(1a) (Věděl jsem, (že (přijde)) a (že (mi (dá pusu))))).

Způsob analýzy je dán nějakou předem danou gramatikou, nějakým *schématem analýzy*, které je návodem, jak analyzovat věty. Musí jít o schéma, které se postupně a inkrementálně doplňuje o případy, které se předtím nevyskytly. Proces je kontinuálně inkrementální a sotva kdy budeme moci tvrdit, že jsme dospěli k úplné gramatice daného jazyka.

Je-li však stromová banka vytvořena, lze z ní automaticky odvodit *frázovou gramatiku*, v níž minimální podstromy interpretujeme jako *nekontextová* pravidla. Taková gramatika je zárodkem *probabilistické* frázové gramatiky, protože jednotlivá pravidla se ve výchozím korpusu vyskytují s určitými četnostmi, které lze považovat za první aproximaci pravděpodobností, s nimiž se taková pravidla mohou vyskytovat v budoucím textu podobného typu.

Pražský závislostní stromový korpus Pro češtinu je nyní k dispozici pražský závislostní stromový korpus (Prague Dependency Tree Bank), vybudovaný skupinou prof. Hajičové na MFF UK a čítající cca 100 000 vět.

0.3.10 Situace v češtině

Závěrem uvedme základní informace o tom, jak vypadá situace pro češtinu. Na podzim roku 1994 byl na FF UK založen *Ústav českého národního korpusu*, v němž se nyní buduje *Český národní korpus – ČNK*. Ke konci roku 1995 byl již k dispozici jeho základ, v němž bylo uloženo cca 30 000 000 slovních tvarů, a na konci r. 1996 již *ČNK* obsahoval téměř 100 mil. českých slovních tvarů. V r. 1999 to už bylo cca 140 mil. a ke konci r. 2000 lze počítat s 200 mil. slovních tvarů. Vedle *ÚČNK* se na této práci podílejí další pracoviště na UK, a to *Ústav teoretické a počítačnické lingvistiky FF UK (ÚTKL)*, *Ústav formální a aplikované lingvistiky MFF UK (ÚFAL)*, dále *Ústav pro jazyk český AV ČR (ÚJČ)* a v neposlední řadě i *Ústav českého jazyka FF MU* a *Katedra informačních technologií* na Fakultě informatiky MU. Na posledně jmenovaném pracovišti vznikla v r. 1997 *Laboratoř zpracování přirozeného jazyka (LZPJ)*, která paralelně buduje a udržuje korpusy českých textů, konkrétně korpus ESO, který v současnosti čítá kolem 160 mil. slovních tvarů, a dále plně gramaticky značkový korpus DESAM v rozsahu něco přes 1 mil. slovních tvarů. Tento korpus na rozdíl od pražských experimentů se stochastickým značkovačem J. Hajiče byl vytvořen převážně manuálně, ovšem míra úspěšnosti značkování v něm nyní dosahuje kolem 98

0.3.11 Struktura ČNK

Popis, přístup Korpusy na FI MU, přístup k nim: PUBL, FIT, DESAM

V květnu 96 byl GA ČR schválen komplexní grantový projekt *Čeština ve věku počítačů* ve výši cca 30 mil. Kč a s dobou trvání 6 let. Nositelkou grantu je prof. E. Hajičová z *Ústavu formální a aplikované lingvistiky MFF UK* a jeho hlavním cílem je:

1. dobudovat Český národní korpus tak, aby ke konci r. 2000 obsahoval cca 200 mil. českých slovních tvarů a byl přístupný pro lingvistickou i ostatní odbornou veřejnost (prostřednictvím Cesnetu a Internetu).
2. gramaticky (slovní druhy a gramatické kategorie) označkovat cca 10 mil. slovních tvarů v rámci ČNK.
3. vytvořit základ stromové banky pro češtinu – min. v rozsahu 1 mil. slovních tvarů (*Pražský závislostní korpus, PDTB*).
4. vytvořit soubor potřebných korpusových nástrojů, tj. programové vybavení zahrnující:

- manažery
 - značkovače – gramatické, syntaktické, sémantické
 - desambiguátory
 - třídící, konkordanční a konverzní programy.
5. přenést SSJČ na počítačová média (skenováním).
 6. rozpracovat přípravu elektronické lexikální datové báze pro češtinu, která se stane východiskem pro budování nového velkého slovníku češtiny (primárně elektronického).

Struktura textů ukládaných do korpusu se vyznačuje analyzátozem SGML. Pro gramatické značkování je v LZPJ k dispozici analyzátor a lematizátor (tagger) AJKA vytvořený v LZPJ na FI MU (Sedláček, Veber, 1999), který je dnes schopen pracovat se 150 000 českých kmenů a dovede každému rozpoznávanému slovnímu tvaru přiřadit jeho slovní druh(y) a odpovídající gramatické kategorie. Na rozdíl od pravděpodobnostně orientovaných analyzátorů pro angličtinu je AJKA založena na úplné pravidlové morfologické analýze češtiny, proti které je podobná analýza angličtiny spíše dětskou hračkou. Podobné lematizující programy existují i pro slovenštinu a ruštinu a dále pro angličtinu, němčinu a francouzštinu (Osolsobě, Ševeček, 1995).

Vedle již uvedených důvodů korpusy potřebujeme i s ohledem na náš budoucí vstup do EU: i když jednacím jazykem je zde do značné míry angličtina, překládání mezi jazyky uvnitř EU je nevyhnutelné. Vznikají proto **paralelní korpusy** využívané při budování systémů strojového překladu a tvorbě vícejazyčných a dnes už primárně elektronických slovníků. Už delší dobu je jasné, že EU počítá s Polskem, Maďarskem a Českou republikou jako prvními východoevropskými členy EU – odráží se to např. v tom, že se uskutečnily některé společné slovníkové projekty, jako např. CEGLEX (Central European Generic Lexicon) zahrnující primárně polštinu, maďarštinu a češtinu a také EuroWordNet 2, jenž vedle šesti západoevropských jazyků obsahoval i češtinu a estonštinu. Český WordNet byl v rámci EuroWordNetu 2 budován právě v LZPJ na půdě FI MU.

0.4 Reprezentace morfologických struktur (pro češtinu)

Morfologie zahrnuje tři oblasti, o nichž je potřeba se zmínit:

- formální morfologii, vlastní tvarosloví – zahrnující flexi, tj. tvoření slovních tvarů ohýbáním, sem patří deklinace, skloňování: substantiv, adjektiv, zájmen a číslovek, konjugace, časování: sloves, stupňování: adjektiv a adverbii,

- **derivační morfologii**, tedy tvoření slov – popisuje odvozování (derivování) nových slov z jiných, obvykle považovaných za základní:
 - den* → *denní* (substantivum – adjektivum)
 - učit* → *učení* (sloveso – substantivum, tzv. deverbativum)
 - vypracovat* → *vypracovaný* (sloveso – deadjektivum)
 - myslet* → *vymyslet, rozmyslet* (sloveso – sloveso)
 - rychlý* → *rychle* (adjektivum – adverbium)
 - rychlý* → *rychlost* (adjektivum – substantivum)
 - student* → *studentka* (subst. – subst.: přechylování)
 - dům* – *domek* – *domeček* (subst. – subst.: tvoření deminutiv)
 - bába* – *babizna* (subst. – subst.: tvoření augmentativ).

Slovotvorba představuje v rámci morfologie samostatnou subdisciplínu a její algoritmický popis je zatím zvládnut jen v základních obrysech. Jednou z prvních věcí potřebných v tomto ohledu pro češtinu je vytvoření tzv. **derivačního slovníku**, tj. slovníku, jehož základními položkami jsou kořeny českých slov plus morfémy, které se s kořeny mohou kombinovat. Předpokládaný počet slovních kořenů není příliš velký, odhadem lze říci, že by se v češtině měl pohybovat kolem 10 000 položek.

- **významosloví** – jinak řečeno teorie slovních druhů. Už zběžný pohled na libovolný text ukazuje, že se v něm vyskytují typy slov seskupující se podle svých vlastností do jednotlivých tříd. Slova v textu či textech lze podle určitých kritérií klasifikovat a získat jednotlivé třídy slov, tedy v obvyklé terminologii **slovní druhy**. Tato klasifikace není triviální a opírá se o kombinaci tří základních kritérií:
 1. o **formu slova**, tj. o způsob jeho flexe, ohýbání, tak dostáváme členění na slova *ohebná* a *neohebná*,
 2. o **význam slova**, podle tohoto kritéria substantiva primárně označují bytosti, různé konkrétní i abstraktní objekty, procesy, události; slovesa mají pak převážně význam *relační*, tj. označují vztahy, vlastnosti (jednomístné vztahy), stavy, děje a činnosti; adjektiva nejčastěji označují vlastnosti objektů označovaných substantivy a adverbia lze významově charakterizovat jako vlastnosti vlastností nebo vlastnosti dějů či činností. Samostatným slovním druhem jsou z hlediska významu *číslovky*, které fungují jako kvantifikátory. Významově prázdnou třídu slov představují *zájmena*, představující svého druhu proměnné, ale právě díky této své vlastnosti je tento slovní druh překvapivě kompaktní. Nemáme ovšem zatím k dispozici seznamy, které by obsahovaly slova klasifikovaná podle svého významu, roztroušeně lze údaje tohoto druhu najít ve slovnících. Pořízení těchto seznamů je jedním z úkolů korpusové lingvistiky a počítačové lexikografie.
 3. o **syntaktickou funkci slova**, tj. o to, jak slovo funguje ve větě, jakou její složkou může být. Obecně to lze říci tak, že některá slova fungují ve větě jako *řídící*

(slovesa a substantiva), jiná jako **modifikující** (adjektiva, adverbia) a jednu skupinu tvoří slova, která můžeme nejlépe charakterizovat **pomocná, funkční, syntaktická** – sem typicky patří **předložky a spojky**. Samostatným slovním druhem jsou po syntaktické stránce **částice, partikule**, které mají nejčastěji povahu celovětných nebo členských **modifikátorů** a při budování algoritmického popisu větné stavby jsou s nimi značné potíže.

Celá klasifikace slovních druhů, s níž budeme nadále pracovat, se opírá o to, co najdeme ve standardních gramatikách (např. MČ II) a obsahuje obvyklých deset slovních druhů plus provizorně zkratky jako samostatnou třídu slov (podrobnější analýza zkratk naznačuje, že většinou mají substantivní povahu a svou vnitřní strukturou představují i dosti složité jmenné skupiny). Ve skutečnosti, jak lze vidět z níže uvedené formální reprezentace zachycující výše zmíněnou klasifikaci, zavádíme ještě uvnitř některých slovních druhů jemnější rozklady, subklasifikace: to platí např. o zájmenech, číslovkách, adverbích a slovesech, ale i o substantivech a třeba spojkách a částicích. Celkově však toto členění nelze pokládat za konečné, a to jak pokud jde o slovní druhy samotné, tak i jejich subklasifikace. Příkladem mohou být podstatná jména, uvnitř nichž v každém případě dále potřebujeme rozlišit vlastní jména a příjmení, geografické názvy a názvy institucí a další – to však v níže uvedené klasifikaci ještě není začleněno.

0.4.1 Přehled notace pro českou morfologii a syntax

Notace je ve shodě se současnými gramatikami a teoretickými hledisky, která se v nich uplatňují (MČ II, Grepl, Karlík, 199?, Jelínek et al., 1995). Je vybudována tak, že jednotlivé gramatické kategorie jsou zavedeny jako atributy, které nabývají možných hodnot. Základními atributy jsou pak slovní druhy, nabývající podle daného slovního druhu hodnot 0-9 (viz níže) a hodnoty X (zkratky). Následuje výčet slovních druhů včetně podtříd a jejich standardních gramatických kategorií. Celkově má notace **otevřený** charakter, tj. lze ji podle potřeby doplňovat a rozšiřovat a zachovat přitom kompatibilitu s předchozím stavem. Současné úpravy představují především zavedení dalších zjemnění a subklasifikací.

Přehled gramatických značek pro:

a) slovní druhy,

b) jejich odpovídající gramatické kategorie:

Princip konstrukce značky je dán následující konvencí: atributy jsou značeny malými písmeny, hodnoty atributů velkými písmeny nebo číslicemi. Značky tedy nejsou atomické objekty, mají svou pravidelnou strukturu, jíž se dále využívá např. v syntaktické analýze.

k1, "subs", substantivum, podstatné jméno: rod=gM -mužský živ.,

gI=mužs.než., gF=ženský, gN=střední
číslo=nS=singulár, nP=plurál
pád=c1,2,3,4,5,6,7

- k2, "adj", adjektivum, přídavné jméno, rod u přivlastňovacích=h, adjektiva rozlišují stejné kategorie jako substantiva, tj.rod=g, číslo=n a pád=c, navíc pak klad=eA, zápor=eN a stupeň d1=pozitiv, d2=komparativ, d3=superlativ
- k3, "pron", pronomen, zájmena se dále člení na osobní=P, ukazovací=D, přivlastňovací -- posesivní=O, vztažná -- relativní=R, tázací=Q, neurčitá=U, zvrtná, reflexivní=X, zájmena rozlišují stejné kategorie jako substantiva, tj.g,n,c, ovšem některá z nich, především osobní, jsou bezrodá
- k4, "num", numeralia, číslovky, rozpadají se dále na základní=O, řadové=C, násobné=M, podílné=D, jinak číslovky nesou stejné kategorie jako substantiva, tj.g,n,c
- k5, "verb", verbum, sloveso, nese kategorie: klad=eA, zápor=eN, osoba=p1,2,3 (první, druhá, třetí), číslo=nS=singulár, nP=plurál, čas=tP=přítomný, tM=minulý, tF=budoucí
způsob=mI=indikativ, mR=imperativ, mC=kondicionál
vid=aP=dokonavý, perfektivní aI=nedokonavý, imperf.
- k6, "adv", adverbium, příslovce, člení se na adv.způsobu=M, času=T, místa=L, modální=D, příčiny=C, typické kategorie: klad=eA, zápor=aN, stupeň=d1, d2, d3
- k7, "prep", prepozice, předložka, rozlišuje pád=c2,3,4,6,7
- k8, "conj", konjunkce, spojka, člení se na souřadící=C a podřadící=S
- k9, "part", partikule, částice, zatím se člení na pravděpodobnostní=P,

rematizační=R, měrové=Q
k0 "intr", interjekce, citoslovce
kX "abbr", zkratky, zkratková slova

Doplňující přehled gramatických kategorií rozlišovaných standardně v češtině:

numerus=číslo=n - "sg"=S, jednotné, "pl"=P, množné
genus=rod (jmenný) "mask anim"=Mn mužs.životný, maskulinum
"mas inan"=In mužs.neživotný,
"fem"=F ženský, femininum
"neu"=N střední, neutrum
U= mužs.než.nebo střední, mužs.živ., neživ.
Y=všechny rody "mask.anim+mask.inan+fem+neu"
kazu=pád=c, "1234567" (1=nominativ, 2=genitiv, 3=dativ, 4=akuzativ,
5=vokativ, 6=lokál, 7=instrumentál)
pers=osoba=p, "1.os=1", "2.os=2", "3.os=3",
stupňování u adjektiv a adverbii=d "1.st-pozitiv", "2.st-komparativ",
"3.st-superlativ"
slovesný způsob, modus=m "indik"=I, indikativ (oznamovací způsob)
"imper"=R, imperativ (rozkazovací způs.)
"kondic"=C, podmiňovací způsob
příčestí, "participium"=part: minulé=M, trpné (mezera)
přechodník, transgresiv=trsg - "prech"=T
čas, temp=t "preteritum"- minulý=M, "prézens"-přítomný=P,
"futurum"-budoucí=F
vid slovesný, aspekt=a "perf"=P, perfektivní, dokonavý
"imperf"=I, imperfektivní, nedokonavý
klad a negace=e, A=kladné - bez ne, ~ s ne
adverbia= "jak"- způsobu=M, "kde"- místa=L, "kdy" - času=T,
"mod"- modální=D, "proč" - příčiny=C, "kolik"=míry=Q
pády u předložek, prep = { "", "2", "3", "4", "6", "7", "4,6", "4,7" };
spojky - conj = "sour"-souřadící, koordinační=C, "podr" -
podřadící, subordinální=S
částice - je připravena subklasifikace rozlišující podle funkce
částice = k9xQ - měrové
= k9xK - kontaktové
= k9xR - rematizátory (omezovací)
= k9xN - navazovací

Typické příklady rozvinuté a zkrácené notace:

k1: tvar "počítač" k: 1 sl.druh: substantivum
g: I rod: muž.neživotný
n: S číslo: singulár
c: 1,4 pád: první nebo čtvrtý
výsledné značky (tags): k1gInSc1, k1gInSc4

k2: tvar adjektiva "rychlý" k: 2 sl.druh: adjektivum
e: A klad (zápor N)
g: M,I rod mužs.živ., muž.neživ.
c: 1,4 pád - nom. nebo akuzativ
d1: stupeň první - pozitiv
výsledné značky: k2eAgMnSc1d1, k1eAgMnSc4d1, k1eAgInSc1d1, k1eAgInSc4d1,

pozn.: adjektiva se shodují se svým řídicím substantivem, u něhož stojí a od něho přebírají tzv.shodové kategorie, tj. g,n,c

k3: tvar osobního zájmena "ty" k: 3 sl.druh: zájmeno, osobní=P
g: nevyjadřuje, tzv.bezrodé
n: S číslo: singulár
c: 1 pád: první, nominativ
výsledná značka: k3xPnSc1

tvar "ty" je však homonymní s tvarem ukazovacího zájmena, jemuž odpovídá značka

k: 3 sl.druh: zájmeno ukazovací
g: M,I rod.mužs.živ.,než.,F žens.,N st
n: P číslo: plurál
c: 1,4 pád (homonymie)
výsledné značky: k3xDgMnPc4, k3xDgInPc1, k3xDgFnPc1, k3xDgNnPc1,
k3xDgInPc4, k3xDgFnPc4, k3xDgNnPc4

pozn.: ukazovací, demonstrativní zájmena se shodují se svým řídicím substantivem, u něhož stojí a od něho přebírají tzv.shodové kategorie, tj. g,n,c

k4: tvar číslovky "tři" k: 4 slovní druh: číslovka

		x: C základní, kardinální
		g: X všechny rody
		n: P číslo: plurál
		c: 1,4,5 pád (homonymie)
výsledná značka:	k4xCgXnPc145	
k5 tvar slovesa "mluvíš"		k: 5 slovní druh: sloveso
		e: A kladný tvar
		p: 2 osoba: druhá
		n: S číslo: singulár
		t: P čas: přítomný
		m: I způsob: indikativ, oznamovací
		a: I vid: imperfektivní, nedokonavý
výsledná značka:	k5eAp2nStPmIaI	
značky pro tvar "mluvil":	k5eApMnStMmPaI, k5eApInStMmPaI	
k6 tvar adverbia "dobře"		k: 6 adverbium, příslovce
značka k6xMeAd1		x: M způsobu, modi
		e: A kladné
		d: 1 pozitiv, první stupeň
k6xTeA	"dnes"	k: 6 adverbium, příslovce
		x: T času, tempori
		e: A kladné
k6xLeA	"tady"	k: 6 adverbium, příslovce
		x: L místa, loci
		e: A kladné
k7 předložka	"na"	k: 7 předložka, prepozice
		c: 4,6 pád
k8 spojka	"že"	k: 8 spojka, konjunkce
		x: S podřadicí, subordinační

k9 částice	"asi"	k: 9 částice, partikule x: P vyjadřuje pravděpodobnost
kX zkratka	"DOS", "NATO"	k: X zkratka, zkratkové slovo

0.4.2 Algoritmický popis (české) morfologie

Algoritmický popis českého tvarosloví, jak jsme už naznačili, zahrnuje deklinaci, konjugaci a stupňování a některé pravidelné derivační (slovotvorné) procesy.

K jeho vytvoření musíme najít způsob, jak formulovat formální pravidla popisující ohýbání slov – ta jsou základem, a jak je potom implementovat. Základní myšlenka spočívá v použití ohýbacích vzorů, jak je známe ze školy, ovšem pro algoritmický popis je nezbytné základní soubor vzorů rozšířit a zjemnit jejich klasifikaci. V klasických mluvnicích se to řeší uvedením výjimek – v algoritmickém popisu se jim můžeme vyhnout, jestliže zavedeme dostatečný počet podvzorů zachycujících příslušné hláskové změny, např. *vlk – vlci, doktor – doktoři, medvídek – medvídko – medvídci, pes – psa, dívka – dívce, den – dne* apod. (podrobně viz Osolsobě, 1994).

Vlastní segmentace pak probíhá zhruba tak, že nejprve se hledají prefixy *nej-* a *ne-*, pak ev.i další – *vy, roz, po, na*. V případě úspěchu se činí pokus najít kmen, pokud je ovšem neúspěšný, odtrhávají se koncovky od konce slova, dokud se kmen nenajde ve slovníku. Díky uložení kmenů a možných pokračování v podobě konečného automatu a struktur trie je hledání velmi rychlé, neboť jeden kmen je společný pro řadu sufixů a intersegmentů.

Na naznačeném algoritmu je založen:

- program LEMMA (Ševeček, 1995) a nověji také AJKA (Sedláček, 1999), které oba rozpoznávají slovní tvary nebo je generují, vstupnímu slovnímu tvaru přiřadí jeho odpovídající gramatické kategorie, tj. slovní druh, pád, číslo, jm. rod (u substantiv, adjektiv, zájmen a číslovek), osobu, čas, číslo, způsob, sl. rod, vid (u sloves) a další u dalších slovních druhů,
- jeho výchozími datovými strukturami jsou vzory (asi 830), kmeny (cca 164 tis.), intersegmenty (cca 460) a koncovkové množiny (počet koncovek cca 127), prefixy (cca 140).

Schéma vytvořené pro LEMMA je v současnosti úspěšně využito pro více jazyků, konkrétně – češtinu, slovenštinu, ruštinu, angličtinu, němčinu, francouzštinu. Základní údaje pro jednotlivé jazyky – počty vzorů, kmenů, velikost slovníku kmenů, jsou uvedeny v tabulce 1:

	Czech	Slovak	Russian	English	German	French
vzory (pocet)	830	488	1150	65	335	325
kmeny (tis.)	165	120	~120	120	130	37
vel.slov.(KB)	660	524	~600	386	665	156
rez.c1 (KB)	25	14	-	10	-	8
rez.c2 (KB)	7	4	-	35	-	55

Tabulka 2 uvádí početní zastoupení slovních druhů v českém slovníku kmenů a počty vzorů u každého slovního druhu.

	cz		vzory		
substantiva	76 400	(1 500 - ista)	376		
slovesa	36 200		180		
adjektiva	43 800		90		
adverbia	1 300		5		
pronomina	137		45	num	32
prepozice	93	spojky	81	partikule	81

– příklad tvaru *s-e-š-i-t-e-m*, ev. *nej-ne-u-věř-i-t-eln-ějš-ímu*, tedy:

1. krok: prefixy, *ne-*, *nej-*, *u-*,

2. krok: prohledávání kmenů, vyčlenění intersegmentů a pak koncovek, uplatnění vzorů a koncovkových množin.

Jednotlivých modifikací morfologického programu LEMMA se užívá např v následujících konkrétních softwarových produktech jako samostatného modulu:

v textových procesorech: – T602, Wintext 3.1, WP 5.1, 6.0, Windows 95 (MS Word v.7 a MS Office), Pragotext, MAT, unixové: Wonder Word a Edit, Ami Pro (jen thesaurus).

Funkce:

– korekce překlepů

– nabídka možných tvarů (s ohledem na typy překlepů a chyb)

- nabídka synonym a antonym (thesaurus)
 - dělení slov – to však dělá samostatný dělicí program
- v sázecích systémech:
- Corell, Quark, TeX: zde se nejvíce se využívá dělení slov
- ve fulltextových aplikacích využívajících lemmatizace, tj. přiřazení základního tvaru k libovolnému vstupnímu – viz též níže vyhledávání informací – sem patří systémy jako TOPIC (firma VERITY, např. hledání ve fulltextové db. právních dokumentů)
- OCR systémy
- v překladových programech a překladových elektronických slovnících příkladem za všechny je oboustranný anglicko-český a německo-český slovník Lingea Lexicon 2 (Ševeček, 1998).

0.5 Re prezentace syntaktických struktur – gramatiky

0.5.1 Gramatiky pro popis PJ

Struktura syntaktického analyzátoru pro daný jazyk je v jistém smyslu popisem tohoto jazyka, ovšem takový popis nebývá obvykle pro lidi příliš transparentní. Často je závislý na konkrétní implementaci a implementace, i když jsou psány v některém z hlavních programovacích jazyků, se mohou od sebe podstatně lišit.

To byl jeden z hlavních důvodů, který vedl badatele k tomu, že se postupně odvraceli od procedurálních definic sémantiky programovacích jazyků a svou pozornost obrátili k popisům deklarativním. Podobné úvahy jsou na místě i u programů pro NLP: to, co potřebujeme, je jak syntakticky, tak i sémanticky spolehlivý popis zpracovávaného přirozeného jazyka (nebo jeho aproximace), máme-li získat rozumnou představu o tom, jak se daný systém bude chovat v rozdílných podmínkách.

Jazyk lze chápat jako **množinu**, členství v níž lze přesně specifikovat konečným souborem pravidel. Množina složených jazykových výrazů není v PJ konečná, takže nelze podat jejich plný výčet. Pokud je v současnosti známo, žádný PJ není konečným jazykem. Okruh konstrukcí, které činí PJ jako čeština nekonečným, je dosti velký. Např. spojka *a* připouští v češtině spojení neomezeného počtu vět a podobně tak i vztažné věty mohou obsahovat slovesné skupiny, které mohou obsahovat jmenné skupiny, které mohou obsahovat vztažné věty, které mohou obsahovat slovesné skupiny, které ...

To, co potřebujeme, jsou tedy **formální** (tj. matematické) systémy, které umožňují definovat členství v nekonečné množině jazykových výrazů a každému členu této množiny přiřadit jeho strukturní popis, a to prostřednictvím konečného souboru pravidel.

Gramatikami tedy budeme rozumět **formální systémy**, které vedle právě zmíněného kritéria splňují ještě tři další:

1. gramatiky jsou vyjádřeny v deklarativním formalismu obsahujícím pouze informaci o tom, které objekty se spolu kombinují a jaké jsou vlastnosti výsledného objektu, tj. tento formalismus neobsahuje žádnou vnější procedurální informaci o tom, jak dát tyto objekty k sobě (taková informace je např. implicitně obsažena v tzv. přechodových sítích).
2. gramatiky v prezentovaném pojetí transparentně spojují každý přípustný řetězec (výraz jazyka) s jeho implicitním strukturním popisem bez nutnosti uvádět explicitní informace pro budování struktur (jak to vyžadují např. ATN).
3. gramatiky přímo specifikují pořadí prvků v řetězu a tudíž se v nich nečiní pokusy rekonstruovat nějaký hypotetický podkladový slovosled.

0.5.2 Gramatika jako reprezentace znalosti

Gramatiky, jak se jimi budeme dále zabývat, jsou deklarativní a z největší části jsou založeny na dekompozici syntaktických kategorií (zhruba slovní druhy) na složky známé jako rasy. Takto pojaté gramatiky podporují kompozicionální přístup k významu, v jehož rámci každý dobře utvořený výraz jazyka má svůj vlastní význam, a to význam složený z významů podvýrazů, které jej tvoří. To je kontext, v němž syntaktická struktura vtisknutá výrazu je klíčovým prvkem pro určení jeho významu.

Z hlediska ZPJ lze zkoumání gramatik pokládat za součást výzkumů v oblasti reprezentace znalosti. Na gramatiku můžeme pohlížet jako na prostředek pro reprezentování jistých znalostí o jazyce, a to natolik explicitně a formálně, že tyto znalosti mohou být dostupné stroji.

V této souvislosti je však třeba zodpovědět několik podstatných otázek:

1. jaký formální systém je pro daný jazyk nejvhodnější, tj. jaký typ jazyka máme před sebou?
2. jaký notační systém zvolit? – toto rozhodování je závislé na přihlédnutí ke kritériím přirozenosti popisu jazyka, matematické síly zvoleného aparátu a výpočetní efektivity.
 - Požadavek přirozenosti vede lingvisty k tomu, aby popis byl formulován přehledně a srozumitelně, byl snadno modifikovatelný a vyjadřoval relevantní generalizace.
 - Poměrně nevelké notační modifikace mohou na jedné straně výrazně omezit třídu vyjádřitelných gramatik a na druhé straně mohou naopak vést k radikálnímu zvýšení potenciální matematické mohutnosti charakterizovaného systému.
 - Formalismus gramatik vytvářený teoretickými lingvisty je obvykle předmětem pozornosti jen pro další teoretické lingvisty. Gramatické formalismy pro počítače musí být podobně jako programovací jazyky srozumitelné jak pro lidi, tak i pro

stroje a navíc zvládnutelné v realistickém čase. Problémy, které vznikají při navrhování gramatických formalismů, jsou vsutku shodné s otázkami, které se objevují při návrzích deklarativních počítačových jazyků pro reprezentaci znalostí.

3. jak deskriptivně adekvátní má daný popis být? – např. jde-li nám o popis naprosto přesný či jen přibližně adekvátní.

Formalismy, k nimž obrátíme svou pozornost v dalším výkladu, budou reprezentovat v podstatě nekontextové frázové gramatiky a budou to gramatiky vymezených klauzulí (DCG) a případně i formalismus PATR (Gazdar, Mellish, 1989).

Všechny druhy gramatik užívaných v počítačové lingvistice využívají v té či oné podobě:

- reprezentaci syntaktických kategorií nebo „slovních druhů“
- datové typy pro slova (slovní formy, tj. slovník)
- datové typy pro syntaktická (morfologická) pravidla
- datové typy pro syntaktické struktury.

Celou gramatiku lze pak chápat jako užití konkrétních datových typů složených z prvních tří jednotek. **Analyzátor** je algoritmus, který bere gramatiku spolu s předloženým řetězem a snaží se vrátit jednu nebo více instancí datového typu syntaktické struktury. Úplný gramatický formalismus tedy poskytuje notaci pro specifikování syntaktických kategorií, slovníkových hesel, gramatických pravidel (ev. i více typů) a syntaktických struktur.

0.5.3 Formální gramatiky

Soubor formálních pravidel, která umožňují generovat nebo rozpoznávat české věty a současně jim přiřazovat popisy jejich struktury, nazveme **formální gramatikou** (přesná definice následuje v dalším oddíle).

Vrátíme se nyní k větě

(v-1) *Ta jeho druhá žena vášnivě miluje rychlá auta.*

Ukázali jsme už, že tato věta se skládá z větných členů, jimiž jsou podmět a přísudek nebo, jinými slovy, lze ji rozčlenit na část podmětovou a část přísudkovou. Jestliže pro větu uijeme označení **S**, pro podmět **Np1** a pro přísudek **Vp**, pak tvrzení, že „větu (v-1) lze rozložit na podmět a přísudek“, můžeme zapsat jako pravidlo:

(p-1) $S \rightarrow Np1 Vp$,

Čtenář si právem může klást otázku, proč jsme nepoužili označení pomocí jiných symbolů, např. **V** pro větu, **Po** pro podmět a **Přís** pro přísudek a tedy i pravidla

(p-1a) $V \rightarrow Po Přís$,

které by rovněž bylo správným zápisem našeho tvrzení.

Je pravda, že neterminální symboly lze volit různě, musí však být splněna jedna podmínka: vztahy mezi prvky věty musí být formulovány tak, aby výsledný popis adekvátně postihoval strukturu věty a byl ve shodě s naší lingvistickou intuicí.

V oddíle (**Použitá symbolika**) definujeme symboliku, která vychází z konvencí zavedených v současných gramatikách češtiny, a opíráme se přitom především o mezinárodní (latinskou) gramatickou terminologii.

Definice gramatik

Od intuitivního vymezení gramatiky **g1** uvedeného výše přejdeme nyní k formální definici, kterou lze najít v literatuře, viz např. práci Češka a Rábová (1985), ale i Chomsky (1966).

Vedle formální definice pojmu gramatiky si připomeneme i klasifikaci gramatik. Je důležité uvědomit si, že tento přístup je **neutrální** vzhledem ke kterémukoli přirozenému jazyku, což znamená, že je také bezprostředně aplikovatelný nejen na češtinu, ale i třeba němčinu, angličtinu nebo francouzštinu a ruštinu a další. Lingvisticky orientovaný výklad uvedené problematiky je v klasické podobě podán u Chomského (1966), což je práce, kterou by si měl přečíst každý adept počítačové lingvistiky. Čtenáři, který se chce dovědět více o formální teorii jazyků a gramatik a vztazích k teorii automatů, doporučujeme věnovat pozornost např. práci Novotného (1988) a také kapitolám Chomského a Millera z knihy *Handbook of Mathematical Psychology* (Chomsky, Miller, 1965).

Gramatika v tomto chápání představuje formální prostředek, pomocí něhož můžeme vymezit jak **konečné**, tak **nekonečné** jazyky, přičemž gramatika sama je **konečná**.

Nejprve uvedeme potřebné výchozí pojmy: Prvním z nich je **abeceda**, jíž rozumíme neprázdnou množinu prvků – symbolů abecedy. Jako příklad lze uvést třeba latinskou abecedu čítající 52 symbolů (velká i malá písmena) nebo českou abecedu, která celkem obsahuje 82 symbolů.

Dalším je **řetězec** (ev. slovo). Řetězcem nad danou abecedou rozumíme nějakou posloupnost symbolů abecedy. Posloupnost, která neobsahuje žádný symbol, nazveme **prázdným řetězcem** a budeme ji značit e .

Přesněji řečeno, řetězec nad abecedou T definujeme takto:

1. prázdný řetězec e je řetězec nad abecedou T ,
2. je-li x řetězec nad T a $a \in T$, pak xa je řetězec nad T ,
3. y je řetězec nad T tehdy a jen tehdy, lze-li y získat aplikací pravidel (1) a (2).

Máme-li řetězce x a y a připojíme-li y za x , vznikne řetězec xy . Této operaci říkáme **zřetězení** (konkatenace).

Je dána abeceda T . Pak T^* je množina všech řetězců nad abecedou T včetně prázdného řetězce a T^+ je množina všech řetězců nad T kromě prázdného řetězce e , tj. $T^* = T^+ \cup \{e\}$. Množinu L , pro niž platí $L \subseteq T^*$ (případně $L \subseteq T^+$, pokud $e \neq L$), nazýváme jazykem nad abecedou T . Jazykem tedy může být libovolná podmnožina řetězců nad danou abecedou.

Budeme pracovat se dvěma disjunktivními abecedami (množinami) symbolů:

1. abecedou N (množiny) neterminálních symbolů, které v popisu jazyka interpretujeme jako syntaktické kategorie,
2. abecedou T (množiny) terminálních symbolů, jež interpretujeme (nejčastěji) jako slova daného jazyka,
3. sjednocení obou abeced N a T , tj. $N \cup T$, nazýváme slovníkem gramatiky.

V dalším výkladu budeme pro zápis terminálních a neterminálních symbolů a z nich tvořených řetězců užívat následující konvence, jíž jsme se ostatně přidržovali již výše:

1. a, b, c, d, \dots – označují terminální symboly
2. A, B, C, D, \dots – označují neterminální symboly
3. U, V, \dots, Z – označují terminální nebo neterminální symboly
4. $\alpha, \beta, \dots, \omega$ – označují řetězce terminálních a neterminálních symbolů
5. u, v, \dots, z – označují řetězce pouze terminálních symbolů

Nyní jsme připraveni definovat formální gramatiku $G1$.

Gramatika $G1$ je uspořádaná čtveřice

$$g1 = \{N, T, P, S\},$$

- kde N je konečná množina neterminálních symbolů, které interpretujeme jako syntaktické kategorie,
- T je množina terminálních symbolů, jež interpretujeme jako konkrétní české slovní tvary, a platí, že $N \cap T = \emptyset$,
- P je konečná podmnožina kartézského součinu $(N \cup T)^* N \quad (N \cup T)^* \times (N \cup T)^*$,
- $S \in N$ je tzv. vyznačený počáteční symbol gramatiky G ,

- prvek (α, β) množiny P nazýváme **přepisovacím pravidlem** a budeme jej zapisovat ve tvaru $\alpha \rightarrow \beta$. Řetězec α nazýváme **levou stranou pravidla**, řetězec β **pravou stranou** přepisovacího pravidla.

Jádrem gramatiky tedy je konečná množina přepisovacích pravidel. Každé pravidlo má tvar uspořádané dvojice (α, β) řetězců a stanovuje možné nahrazení řetězce α řetězcem β . Řetězec α obsahuje alespoň jeden neterminální symbol, řetězec β je prvek sjednocení $(N \cup T^*)$.

Nechť λ a μ jsou řetězce z $(N \cup T)^*$. Pak mezi nimi platí relace \xRightarrow{G} , která se nazývá **přímá derivace**, jestliže řetězce λ a μ můžeme zapsat ve tvaru

$$\begin{aligned}\lambda &= \gamma\alpha\delta \\ \mu &= \gamma\beta\delta,\end{aligned}$$

kde γ a δ jsou libovolné řetězce z $(N \cup T)^*$ a $\alpha \rightarrow \beta$ je nějaké přepisovací pravidlo.

Dojdeme-li v posloupnosti přímých derivací k řetězci, který obsahuje pouze terminální symboly, pak již nelze aplikovat žádné přepisovací pravidlo a proces generování končí. Z této skutečnosti, která plyne z definice pravidla, je odvozen název množiny T jako množiny **terminálních symbolů**.

Jestliže existuje posloupnost přímých derivací $\nu_{i-1} \Rightarrow \nu_i, i = 1, \dots, n, n > 1$ taková, že platí: $\lambda = \nu_0 \Rightarrow \nu_1 \Rightarrow \dots \Rightarrow \nu_{n-1} \Rightarrow \nu = \mu$, nazýváme ji **derivace** a značíme ji $\xRightarrow{+}$. Tuto posloupnost nazýváme **derivací délky n** .

Jestliže v gramatice G platí pro řetězce λ a μ relace $\lambda \xRightarrow{+} \mu$ nebo identita $\lambda = \mu$, pak píšeme $\lambda \xRightarrow{*} \mu$. Relace $\xRightarrow{*}$ je tranzitivním a reflexivním uzávěrem relace **přímé derivace**.

Důležitým prostředkem pro grafické vyjádření struktury věty (její derivace) je graf-strom, který se nazývá **derivační** nebo **syntaktický strom** věty. Přesněji řečeno, strom je **orientovaný acyklický graf** s následujícími vlastnostmi:

1. existuje jediný uzel, tzv. **kořen** stromu, do něhož nevstupuje žádná hrana,
2. do všech ostatních uzlů vstupuje právě jedna hrana,
3. uzly, z nich žádná hrana nevystupuje, se nazývají **koncové (terminální)** nebo také **listy**,
4. při kreslení se zachovává konvence, že kořen je nejvýše a všechny hrany jsou orientovány směrem dolů,
5. uspořádání hran zachovává **slovoslednou relaci**, tj. pořadí slov ve větě (zleva doprava).

Je-li G gramatika, pak řetězec $\alpha \in (N \cup T)^*$ se nazývá **větná forma** právě tehdy, když platí $S \xRightarrow{*} \alpha$, tj. řetězec α je generovatelný z počátečního symbolu S . Větná forma, která obsahuje pouze terminální symboly, se nazývá **věta**. Jazyk $L(G)$ generovaný gramatikou G je definován množinou všech vět:

$$L(G) = \{w \mid S \xRightarrow{*} w \wedge w \in T^*\}.$$

Množinu vět generovaných gramatikou nazýváme **jazyk** a dále rozlišujeme **slabou generativní kapacitu gramatiky**, jíž je jazyk $L(G)$ (množina všech vět generovaných gramatikou G), který je gramatika G schopna generovat, a **silnou generativní kapacitu** – což je množina **syntaktických stromů** (strukturních popisů) přiřazovaných větám jazyka L generovaného gramatikou G .

0.5.4 Typy gramatik

Gramatiky lze klasifikovat do typů podle tvaru přepisovacích pravidel. Je obvyklé vymezovat čtyři typy gramatik, které se nazývají **typ 0**, **typ 1**, **typ 2** a **typ 3**.

Typ 0

Gramatika typu 0 obsahuje pravidla v nejobecnějším tvaru, kdy platí

$$\alpha \rightarrow \beta, \alpha \in (N \cup T)^* N \quad (N \cup T)^*, \beta \in (N \cup T)^*.$$

Protože se neklade žádné omezení na tvar pravidel a povoluje se přepisovat řetězce na řetězce, mluvíme také o **neomezených přepisovacích systémech**.

Typ 1

Gramatika typu 1 obsahuje pravidla tvaru

$\alpha A \beta \rightarrow \alpha \gamma \beta$, $A \in N$, $\alpha, \beta \in (N \cup T)^*$, $\gamma \in (N \cup T)^+$ nebo $S \rightarrow e$.

Gramatiky typu 1 se také nazývají **gramatikami kontextovými**, protože v kontextových pravidlech lze neterminální symbol A nahradit řetězcem γ pouze tehdy, je-li jeho pravým kontextem řetězec β a levým kontextem řetězec α .

Kontextové gramatiky neobsahují pravidla tvaru $\alpha A \beta \rightarrow \alpha \beta$, a tedy nepřipouštějí, aby neterminální symbol byl nahrazen prázdným řetězcem. Jinými slovy, při generování věty nemůže dojít ke zkracování generovaných řetězců.

Typ 2

Gramatika typu 2 obsahuje pravidla tvaru

$A \rightarrow \gamma$, $A \in N$, $\gamma \in (N \cup T)^*$.

Nazýváme je také gramatikami **nekontextovými**, protože nahrazení neterminálního symbolu A na levé straně pravidla řetězcem γ lze provést bez ohledu na jakékoli okolí, v němž by se neterminální symbol A mohl vyskytovat.

Pro popis syntaktické stavby přirozených jazyků jsou nejzajímavější právě nekontextové gramatiky. Gramatika **g1** popsaná výše je příkladem nekontextové gramatiky pro češtinu. Podobně gramatiky vymezených klauzulí v PROLOGU, o nichž bude řeč níže, vycházejí z formalismu nekontextových gramatik.

Typ 3

Gramatika typu 3 je tvořena pravidly ve tvaru

$A \rightarrow xB$ nebo $A \rightarrow x$; $A, B \in N$, $x \in T^*$.

Protože jediný možný neterminální symbol na pravé straně pravidla stojí zcela vpravo, mluvíme také o **pravé lineární gramatice**. Poznamenejme ještě, že gramatiky typu 3 se také nazývají **regulárními gramatikami**.

Pro práci s přirozenými jazyky, jak jsme prakticky ukázali výše, zůstávají východiskem gramatiky nekontextové. V lingvistické literatuře posledních 20-30 let se sice spotřebovalo mnoho papíru na argumenty, které si kladly za cíl ukázat, že nekontextové gramatiky jsou pro popis přirozených jazyků nedostačující a že je potřeba zavést gramatiky silnější – **transformační** (viz již Chomsky, 1957), poslední práce (např. Gazdar, 1982, Gazdar, Mellish, 1989, Pereira, 1983) však obsahují jejich určitou rehabilitaci. Zejména se podařilo ukázat, že implementace nekontextových gramatik v PROLOGU v podobě tzv. **gramatik vymezených klauzulí** (**definite clause grammars = DCG**), o nichž bude vzápětí řeč, umožňuje zachovat nekontextovou podobu pravidel a současně získat kontextovou citlivost tak potřebnou pro formální popis gramatické shody a dalších kontextově podmíněných

gramatických jevů v přirozených jazycích.

0.5.5 Několik slov o PROLOGu

Standardním nástrojem v oblasti ZPJ je programovací jazyk PROLOG, který umožňuje poměrně snadno vyjadřovat algoritmy užívané v počítačové lingvistice. Potřebujeme tu často manipulovat se symboly (slovy, morfémy, slovními druhy, různými druhy rysů) a strukturovanými objekty (seznamy, posloupnosti, stromy, grafy), které tyto symboly obsahují.

PROLOG je jazyk vysoké úrovně, v němž lze přímo vyjadřovat operace na symbolech (reprezentovaných jako atomy, řetězcy a čísla) a strukturách (reprezentovaných jako seznamy a termy), aniž se musíme starat o to, jak jsou tyto koncepty vyšší úrovně skutečně reprezentovány v počítači. PROLOG umožňuje přesně specifikovat komplexní struktury v termínech abstraktních vzorců (schémat). Rovněž dovoluje prezentovat informace na značně abstraktní úrovni v termínech souboru faktů a vyjadřovat libovolně složité inference.

V ZPJ hraje základní roli koncept *rekurze*. Jazykové objekty jsou popisovány rekurzivními datovými strukturami a operace na těchto rekurzivních strukturách jsou přirozeně formulovány jako *rekurzivní algoritmy*. Podobně jako jiné vyšší programovací jazyky ani PROLOG neomezuje volání predikátových definic (funkcí) sebou samými (přímo nebo nepřímě), takže rekurzivní algoritmy lze v PROLOGU vyjadřovat přímo.

0.5.6 Gramatiky v PROLOGU

V následujícím ukážeme, jak lze přepsat výše uvedenou gramatiku **g1** tak, aby s ní bylo možno pracovat jako s gramatikou v PROLOGU. Nekontextovým gramatikám, jako je **g1**, v PROLOGU odpovídají gramatiky vymezených klauzulí – DC gramatiky.

0.5.7 Nekontextové gramatiky a DC gramatiky

Gramatická pravidla DC gramatiky jsou velmi podobná pravidlům **g1**, mají stejně jako ona levou a pravou stranu a operátor \rightarrow . Podstatný rozdíl je však v tom, že jednotlivé neterminální symboly v **g1** musí být v DC gramatice zapsány jako *predikáty* s příslušným počtem *argumentů*.

Nekontextovou gramatiku **g1** přepíšeme tedy jako DC gramatiku se jménem **g1.pl**, tj. jako textový soubor s tímto jménem. Soubor vytvoříme pomocí některého z již zmíněných editorů, nejlépe NE nebo CSED. S výhodou lze též použít editoru vestavěného v ARITY PROLOGU v. 5.1.

Při přepisování budeme dodržovat tyto konvence:

1. výraz označující konstantu v PROLOGU musí začínat malým písmenem,
2. výraz označující proměnnou musí začínat velkým písmenem,
3. za každým pravidlem píšeme tečku,
4. /* tento text */ jsou pro PROLOG závorky, do nichž umisťujeme poznámky nebo údaje, které potřebujeme jen my sami, a PROLOG je ignoruje. To se týká např. číslování pravidel gramatiky nebo hlaviček oddělujících vlastní pravidla gramatiky od pravidel definujících slovník (viz níže).

Poznamenáváme, že očíslování pravidel v nekontextové gramatice **g1** a v DC gramatice **g1.pl** je shodné, takže čtenář může porovnávat snadno podobu pravidel v **g1** a v **g1.pl**. Princip přepisu pravidel z nekontextové gramatiky do DC gramatiky je následující:

Vyjděme z pravidla gramatiky **g1**

(p-1) $S \rightarrow Np1 Vp$,

jež, jak víme, rozkládá větu na jmennou skupinu v nominativu a slovesnou skupinu, což je vyjádřeno příslušnými neterminálními symboly. V DC gramatice nemůžeme použít jednoduchých neterminálních symbolů jako v **g1**, ale musíme je nahradit příslušnými predikáty. Místo **S** budeme mít v **g1.pl** predikát $s(s(Np1, Vp))$, který má tři argumenty: z nichž dva jsou pro nás nedostupné a také v rámci DC gramatiky neviditelné a jeden – $s(Np1, Vp)$ – zajišťuje vytvoření podstromu definovaného pravidlem (p-1) v grafu-stromu generované nebo rozpoznávané věty – (v-1). Predikát (neterminál) **s** je splněn, jsou-li splněny predikáty odpovídající neterminálům na pravé straně pravidla (p-1):

NP1 tedy odpovídá $np1(Np1)$ a **VP** odpovídá $vp(Vp)$, takže (p-1) odpovídá

/*p-1*/ $s(s(Np1, Vp)) \rightarrow np1(Np1), vp(Vp)$.

Predikáty $np1$ a $vp1$ jsou stejně jako predikát **s** tříargumentové. Podobně budeme postupovat i u dalších pravidel gramatiky **g1**.

Nyní již můžeme uvést přepis pravidel **g1** do pravidel DC gramatiky:

```
/* gramatika g1.pl */
```

```
/*p-1*/    s(s(Np1,Vp))      →    np1(Np1), vp(Vp) .
/*p-2*/    np1(np1(N1))      →    n1(N1) .
/*p-2a*/   np1(np1(Pnd1,N1)) →    pnd1(Pnd1), n1(N1) .
/*p-2b*/   np1(np1(A1,Np1)) →    a1(A1), np1(Np1) .
/*p-2c*/   np1(np1,(Pos1,N1)) →    pos1(Pos1), n1(N1) .
/*p-2d*/   np1(np1,(Num1,N1)) →    num1(Num1), n1(N1) .
/*p-3*/    vp(vp(Adgm,V3,Np4)) →    adgm(Adgm), v3(V3), np4(Np4) .
```

/*p-3a*/	vp(vp(V3,Np4))	→	v3(V3), np4(Np4).
/*p-3b*/	vp(vp(Adgm,V3))	→	adgm(Adgm), v3(V3).
/*p-3c*/	vp(vp(V3))	→	v3(V3).
/*p-4*/	adgm(adgm(Adm))	→	adm(Adm).
/*p-5*/	np4(np4(A4,N4))	→	a4(A4), n4(N4).

/* slovník */

/*p-6*/	pnd1(pnd1(ta))	→	[ta].
/*p-7*/	pos1(pos1(jeho))	→	[jeho].
	pos1(pos1(moje))	→	[moje].
/*p-8*/	num1(num1(první))	→	[první].
	num1(num1(druhá))	→	[druhá].
/*p-9*/	n1(n1(žena))	→	[žena].
	n1(n1(babička))	→	[babička].
/*p-10*/	v3(v3(miluje))	→	[miluje].
	v3(v3(nenávidí))	→	[nenávidí].
/*p-11*/	a1(a1(krásná))	→	[krásná].
	a1(a1(chytrá))	→	[chytrá].
/*p-12*/	a4(a4(rychlá))	→	[rychlá].
	a4(a4(silná))	→	[silná].
/*p-13*/	n4(n4(auta))	→	[auta].
	n4(n4(kuřata))	→	[kuřata].
/*p-14*/	adm(adm(vášnivě))	→	[vášnivě].
	adm(adm(bláznivě))	→	[bláznivě].

Čtenář si jistě povšimne, že proti **g1** obsahuje **g1.p1** několik pravidel navíc. Jejich užití lze snadno vyzkoušet, a tak si ověřit, v čem rozšiřují výchozí nekontextovou gramatiku **g1**. V cvičeních, která následují, je podrobněji naznačen efekt způsobený např. pravidlem /*p-2b*/, které má tu vlastnost, že neterminální symbol (predikát) **np1** se v něm vyskytuje na levé i pravé straně. Takové pravidlo se nazývá *rekurzivní*. Jsou to právě *rekurzivní* pravidla, která umožňují postihnout opakující se pravidelnosti syntaktických struktur přirozeného jazyka a díky nimž, jak se čtenář může snadno sám přesvědčit, může *konečná* gramatika (tj. gramatika s konečným počtem pravidel) generovat *nekonečně* mnoho vět.

0.5.8 Valenční rámce a jejich začlenění do formálních gramatik

V oblasti počítačového zpracování češtiny se nelze obejít bez rozsáhlejšího seznamu českých sloves s jejich valencemi, který by obsahoval pokud možno všechna běžná česká

slovesa a měl také dostatečně formální podobu. Při jeho sestavování jsme mohli opřít o existující počítačový slovník českých kmenů, který je jádrem automatického morfologického analyzátoru a současně lemmatizátoru LEMMA (Osolsobě 1996, Ševeček 1995) Tento slovník v současnosti obsahuje cca 36 tisíc slovesných kmenů, posloužil jako vhodné východisko k pokusu o vytvoření základního valenčního slovníku zahrnujícího v současnosti kolem 15 tisíc českých sloves. Jako další zdroj posloužil díky své elektronické podobě i *Slovník českých synonym* (Pala, Všianský 1995). Výsledkem je tedy Valenční slovník českých sloves, který u vybraných sloves obsahuje i základní frazeologická spojení a některé kolokace. Takto lze získat přirozené východisko též pro vytvoření základního seznamu valencí i u českých substantiv a adjektiv: takový seznam představuje další chybějící článek formálního gramatického popisu češtiny a je nezbytným předpokladem jejího realistického počítačového zpracování.

Jsme si přirozeně vědomi, že dostatečně reprezentativní seznamy českých sloves užívaných v současné češtině budeme moci získat teprve z právě vznikajícího Českého národního korpusu (ČNK) i spolu s jejich frekvenčními charakteristikami. To ale bude vyžadovat ještě určitý čas (odhadem kolem 2 let) a navíc důležitou podmínkou, která musí být splněna, abychom dostali přesnější obraz o distribuci slovních druhů včetně sloves v současné češtině, je gramatické označkování dostatečně velké části ČNK. V současnosti je k dispozici jen korpus DESAM na FI MU, který je pro tento účel s rozsahem cca 1 mil. slovních tvarů nedostačující, resp. může sloužit jen jako základní východisko.

Na rozdíl od seznamu vytvořeného pod vedením N. Svozilové v ÚJČ (Svozilová et al, 1998?), který jednoznačně předpokládá uživatele – člověka, je VSČS primárně orientován na algoritmický popis české syntaxe a její počítačové zpracování – je proto zachycen pomocí formální notace. Abychom mohli dostatečně přesně zachytit české valence, navrhli jsme notační prostředky, které zachycují jak jednotlivé jednoduché valence, tak i jejich možné kombinace mající pak podobu konkrétních valenčních vzorců. Návrh notace valenčních vzorců svým způsobem navazuje na existující strojový slovník českých kmenů a algoritmický popis české morfologie (Osolsobě, 1996). Principy notace pro valenční vzorce jsou uvedeny a objasněny níže v odd.??

Celkově byl materiál pro VSČS byl získán z následujících zdrojů:

1. Slovník českých synonym, NLN, Praha 1995,
2. Slovník spisovné češtiny, Academia, Praha 1994, 2.vyd.
3. počítačový slovník českých kmenů s celkovým rozsahem cca 164 000 jednotek (prefigovaná slovesa a pravidelně tvořená deverbativa, adjektiva a adverbia jsou však v tomto slovníku generována automaticky, takže skutečný rozsah tohoto slovníku je větší než 300 000 položek).

Výchozí soubor získaný z uvedených zdrojů čítal kolem 10 000 tisíc českých sloves. Po jeho zpracování a postupném porovnání se SSJČ jsme dospěli k první verzi seznamu obsahujícímu cca 12 000 českých sloves, který byl ještě doplněn o slovesa získaná z korpusu

DESAM na rozsah cca 15 000 položek, což je rozsah, který lze z hlediska současných potřeb pokládat za dostačující.

Výchozí pojmy

Ve shodě s Čermákem a Holubem (1991), jako výchozí koncept může sloužit kolokabilita, tj. **obecná schopnost slova (a dalších jednotek) spojovat se v textu s jinými**. S tímto termínem by se pravděpodobně dalo vystačit, u sloves je však obvyklé mluvit o valenci zejména proto, že ji lze vyjadřovat morfologickými prostředky, tj. pády. Tuto schopnost sloves vázat na sebe gramaticky ostatní slova můžeme symbolicky reprezentovat v termínech slovních druhů – substantiv a zájmen nebo pomocí specifických pronominálních výrazů jako *koho, co, čeho, komu, čemu, ...*

V literatuře se často diskutuje o tom, zda valence je jevem primárně syntaktickým nebo sémantickým. Většinou se setkáváme s názorem, že valence je záležitostí roviny syntaktické, což se primárně opírá o fakt, že je (v češtině) vyjadřována morfologickými prostředky – pády, které se vazebně pojí s jednotlivými slovesy. Tyto formální prostředky realizace valence by nám však neměly zakrýt podstatu věci, totiž skutečnost, že schopnost slova kombinovat se v textu s jinými slovy je primárně dána sémanticky, tj. významy spojujících se slov. Považujeme proto slovesnou valenci za jev primárně sémantický a chápeme ji jako významem determinovanou schopnost slova kombinovat se s jinými slovy. V dalším budeme usilovat o co nejuplněnější významovou charakterizaci jednotlivých argumentů, i když v dané verzi slovníku zatím pracujeme zatím především s povrchovými pádovými příznaky. Je však jasné, že bez přihlížení k významu sloves nemůžeme rozumně vysvětlit kontextové elipsy typu

(v1) *Otec čte dětem před spaním.* (pohádka)

nebo významově blízké případy – synonyma jako

(v2a) *Matka mluví s otcem o těch penězích.*

(v2b) *Matka vykládá otcovi o těch penězích.*

i když jejich valenční vlastnosti se povrchově liší (viz např. Leech, 1981).

Typy valencí

Při popisu valenci a pak i sestavování slovníku je nejobtížnější vyrovnat se s obligatorností a fakultativností jednotlivých argumentů u sloves, resp. klasifikovat je vzhledem k těmto kritériím. Obtíže, které tu vznikají, však podle našeho názoru signalizují, že čistě syntaktická kritéria obligatornosti a fakultativnosti nejsou dostatečná a často nemožňují dospívat ke konzistentním rozhodnutím. Vzhledem k rozsahu materiálu jsme proto zatím rezignovali na striktní rozlišení obligatorních a fakultativních doplnění a zaměřujeme se primárně na jejich zachycení hlavně v souladu s významem toho kterého predikátu. Jinak řečeno, jde nám hlavně o to, abychom na prvním místě zachytili, co k danému slovesu patří, a teprve na druhém, jak to k němu patří.

V tomto ohledu nejde ovšem jen o naši zkušenost, a proto např. ve shodě se Somersem (Somers 1987) jsme se pokusili rozlišit následujících šest stupňů valenční vázanosti, pro něž zavádíme po řadě i příslušné notační konvence:

1. *integrální, lexikálně determinovaná, nevypustitelná* doplnění, nepřipouštějící substituce argumentů blízkými synonymy a modifikace adjektivy, nevstupují do syntagmatických substitučních paradigmat (pronominalizace), frazeologická spojení, ev. idiomy, frazémy, např. *držet krok, hubu, mít šanci, mít koho—co po ruce* apod. Vyznačujeme je samostatným symbolem #, který signalizuje, že ve skutečnosti jde o samostatnou jednotku. Výše uvedené *držet hubu* a podobně i *držet na koho* zapíšeme tedy jako:

držet
hubu
<na koho>

2. *obligatorní, nutná* doplnění mající pravidelně formu přímých a předložkových pádů a vedlejších vět uvozených např. spojkami *že, aby*. Obligatorní přímé i předložkové pády jsou vyznačeny symbolem & a větná doplnění symbolem \$. Můžeme tedy mít:

rozkázat t
= komu & co
= \$(aby, co) = inf
dokázat t
= komu & co
= \$(že)

3. *fakultativní, nepovinná* doplnění formálně realizovaná podobně jako v předchozím případě přímými a předložkovými pády. Vyskytují se s příslušnými slovesy pravděpodobně dosti často, ale jejich vypustitelnost nepochybně svědčí o jejich fakultativnosti. K označení těchto případů užíváme ?, takže můžeme mít např.:

dopisovat si
= s kým ? o čem,
kde lze mít jak *dopisovat si* bez doplnění, tak i *dopisovat si s kým, dopisovat si o čem* a nakonec rovněž *dopisovat si s kým o čem*.

Zařazujeme sem i případy jako

vyprovázet, vyprovodit t
= koho ? (z čeho, do čeho, na co),

v nichž první doplnění pokládáme za obligatorní, ale ostatní uvedená v závorce mohou a nemusí být přítomna. Lze namítnout, že mají adverbialní povahu (označují místo), je však třeba si uvědomit, že se pojí se slovesem pohybu, u nichž lokální modifikátory nemají podle našeho názoru povahu naprosto volných doplnění.

4. *střední* doplnění – široce determinovaná sémantickou třídou (významem) slovesa. Nejtypičtějším představitelem tohoto typu doplnění, jak ukazují naše data, je ob-

vykle přímý instrumentál s širokým nástrojovým významem. Dále sem mohou patřit i výrazy časové, místní a způsobové, pokud jsou široce predikovány významem odpovídajících sloves – to platí zejména pro doplnění [jak]. Doplnění tohoto typu jsou notačně zachycována pomocí hranatých závorek, např.:

dopovat t

= koho [čím]

nebo

dosáhnout t

= čeho, co [čím], [jak].

Příkladem široce chápaného lokálního modifikátoru tohoto typu může být třeba

dopít, dopíjet t

= co [z čeho] ,

kde význam slovesa implicitně předpokládá doplnění typu "nádoby", které ovšem může a nemusí být přítomno, bývá však přítomno "obvykle".

5. volná doplnění – sem řadíme zcela volná doplnění časová, místní a způsobová určená spojitelná s každým normálním slovesem. Jsou volná do té míry, že nijak sémanticky nevyplývají z významu slovesa, proto je u sloves v současné verzi slovníku nijak nevyznačujeme a předpokládáme, že jsou v případě potřeby doplnitelná. Pokud jsou vyjadřována čistými adverbii jako *doma*, *tady* nebo *ted*, *dnes* nepředstavuje jejich rozpoznání nebo vygenerování zvláštní potíže, složitější je situace u adverbialních předložkových pádů. Zde počítáme se zavedením vhodných sémantických rysů, které mohou pomoci indikovat, že např. *na ulici* ve spojení *plakala na ulici* má povahu volného lokálního modifikátoru. Po dopracování slovníku valencí do definitivní podoby počítáme s vyznačením volných doplnění u jednotlivých sloves pomocí speciálních rysů, které mohou v explicitní podobě vypadat např. takto:

platit t

= komu & co [čím] [za co] <kdy=dnes> <kde=v obchodě, jak=hotově>

6. periferní doplnění, k nimž nepochybně patří částice různého typu, zejména pak částice mající hlavně pragmatickou povahu. Zatím zvolený způsob značení je v daném okamžiku celkem arbitrární a definitivně bude řešen až v průběhu času. Jako příklad uvedme třeba

poslat

= komu ? co |<asi, možná, patrně>|

Je vidět že použitá škála vede ke klasifikaci, která je širší než klasifikace obvyklé v českých gramatikách. Za její přednost pokládáme právě to, že umožňuje zachytit v jednom rámci jak frazeologická spojení na straně jedné, tak případně i různé typy partikulí zejména pragmatické povahy na straně druhé. Mezi nimi se pak nacházejí jednotlivá doplnění počínaje obligatorními až po volné.

Celkově tedy zachycujeme valenci českých sloves v popisovaném slovníku tak, že u každého slovesa uvádíme s ohledem na jeho význam jednotlivé přímé nebo předložkové pády, s nimiž se dané sloveso pojí. Jak lze vidět z příkladů uvedených výše, notačním prostředkem vyznačujícím jednotlivé valenční vzorce je = (také bychom mohli říci "významy", uvozovkami pak chceme naznačit, že ne vždy musí jít o významy, které by přesně odpovídaly slovníkovému členění třeba v SSČ, lze však očekávat, že míra shody bude dosti velká). Základní údaje v tomto ohledu poskytuje tab. 4 níže.

0.5.9 Vztah mezi slovesnými významy a valencemi

Použitelný popis slovesných valencí se neobejde bez jasného rozlišení vztahů mezi jednotlivými valenčními vzorci a slovesnými významy. Nejprve je tedy potřeba mít u jednotlivých sloves k dispozici jejich významy, což může být zachyceno podobně jako ve WN 1.5. Mějme např. sloveso *rovnat* (značka =1, obecně "=n" značí číslo významu v češtině, značka :1, obecně ":n" označuje číslo odpovídajícího ekvivalentního významu ve WN 1.5 (*eqsynonym*)):

rovnat

=1

vyrovnávat, činit rovným, planýrovat hřiště
#+ level:8

=2

urovnávat, uhlazovat vlasy
#+ arrange:1

=3

vyhlazovat látku, povrch
#+ smooth:3

=4

stavět do hranice (dřevo)
#+ stock:6

=5

pořádat, třídit knihy
#+ sort:5

V takto uvedených datech ovšem chybí údaje o valencích: jestliže je doplníme, budou údaje pro sloveso *rovnat* v naší notaci vypadat takto – (k5 značí aktuální slovní druh – zde sloveso, zájmenné výrazy se symbolem V uprostřed tvoří konkrétní valenční vzorec s příslušnými pády charakteristickými pro dané sloveso a daný význam): *rovnat*

=1 (k5 kdo V co)

vyrovnávat, činit rovným, planýrovat hřiště
#+ level:8

```

=2 (k5 kdo V co komu)
  ## urovnávat, uhlazovat vlasy
  #+ arrange:1
=3 (k5 kdo V co)
  ## vyhlazovat látku, povrch
  #+ smooth:3
=4 (k5 kdo V co do čeho)
  ## stavět do hranice (dřevo)
  #+ stock:6
=5 (k5 kdo V co kde)
  ## pořádat, třídit knihy
  #+ sort:5

```

Posledním údajem, který potřebujeme u sloves mít, je informace o sémantické povaze jednotlivých slovesných participantů, které jsou v dosavadní podobě charakterizovány jen příslušnými zájmennými proměnnými. Sémantickou povahou participantů míníme jejich zařazení pod kategorie typu sémantických pádů jako *Agens*, *Patiens*, *Adresát* a další: to lze vhodně provést využitím tzv. vnitřně jazykových vztahů (*Internal Language Relations – ILI*), jak jsou definovány ve *EurowordNetu-1 a 2* (Vossen, 1998).

V dané verzi slovníku pracujeme jen s pádovými příznaky vyznačenými pomocí pronominálních výrazů jako *koho*, *co*, *čeho*, *komu*, *čemu*, ..., které jsou jednak vhodné mnemotechnicky a jednak umožňují pohodlně rozlišovat opozici *životnost* : *neživotnost*. Lze však vidět, že uvedené příznaky je možno v případě potřeby celkem snadno konvertovat do jiné vhodné notace (Horák, ???), která se Subjektové argumenty jsou v dané verzi implicitní a samostatně jsme nezpracovávali ani aritu sloves (predikátů) tak, že bychom každému predikátu přiřazovali aritu pro jeho jednotlivé významy např. ve formě čísla umístěného před rovnítkem vyznačujícím jednotlivé valenční vzorce:

```

adresovat t
3= komu & co
3= co ? na koho|co

```

Je ostatně patrné, že aritu lze z uvedené notace pohodlně odvodit, aniž bychom ji uváděli samostatným číslem. Příznaky jako *jak* a *kolik* uvádíme jen u sloves, u nichž mají v závislosti na významu jednoznačně valenční charakter. Podobně, jak jsme už naznačili, zacházíme i s adverbiálními pády jako *na čem*, *v čem*, *do čeho*, *z čeho*, ..., které uvádíme jen tam, kde jsou podmíněny významem slovesa – tak je tomu zjevně u sloves pohybu.

Příznaky typu *kam*, *kudy*, *kde*, *kdy*, ... v dané verzi u jednotlivých sloves neuvádíme a situaci kolem adverbiálních argumentů budeme řešit rozvinutím (generováním a rozpoznáním) adverbiálních doplnění s významem místa, času a případně i způsobu. Tento krok je založen na teoretickém předpokladu, že uvedená doplnění se obvykle mohou pojit se všemi běžnými slovesy (mimo např. některá slovesa pohybu apod.).

Počítáme tu však s empirickým ověřováním tohoto běžně vyslovovaného předpokladu a porovnáváním s korpusovými daty.

Východiska pro třídy sloves

Popsaný seznam čítající téměř 12 tisíc českých sloves může posloužit jako východisko k vytvoření řady slovesných tříd získaných na základě pádových příznaků (a jejich kombinací), s nimiž se jednotlivá slovesa pojí. Díky celkové složitosti notace a velkému počtu různých valenčních vzorců představuje třídění sloves s jejich valenčními vzorci dosti komplikovaný úkol, pro jehož úplné vyřešení je třeba napsat samostatný program a také v postupných krocích testovat konzistenci zápisu valencí v aktuální verzi valenčního seznamu. Jde o natolik komplexní věc, že zde můžeme zatím nabídnout jen svého druhu sondu poskytující jen zatím získané údaje o základních typech valencí a valenčních vzorců.

Učinili jsme zatím první pokus a pomocí valencí jsme se pokusili vytřídit z našeho seznamu slovesa **pohybu**. Použili jsme k tomu valencí *do čeho* a *z čeho*, které lze považovat za spolehlivé signály místního doplnění. Takto získaný podseznam jsme ještě prošli manuálně a vyřadili některá slovesa, jež se sice vyskytují s valencí *do čeho*, ovšem označují velmi specifickou variantu pohybu jako např. *bít*, *bouchat do čeho*: výsledkem je seznam sloves pohybu, který zatím čítá cca 1700 sloves (z cca 12 tis. sloves). Tento seznam se ještě zjevně rozpadne na menší a sémanticky kompaktnější skupiny podle jednotlivých typů pohybu – k tomuto jemnějšímu třídění použijeme dalších valencí vyskytujících se u sloves v seznamu jako např. *na co* a *v čem* a dalších.

Již získaná data tedy jasně naznačují, že pomocí valenčních vzorců bude možno získat širší sémantickou klasifikaci českých sloves, která bude velmi užitečná pro různé softwarové aplikace.

První soubor údajů se týká **tranzitivity** a **intranztivity**: chápeme je celkem formálně tak, že za **tranzitiva** pokládáme všechna slovesa, která mají ve svém valenčním vzorci akuzativ – i v kombinaci s jinými pády, zatímco mezi **intranztiva** řadíme ta slovesa, u nichž se akuzativ nevyskytuje. Počítáme tu i s případy, kdy je sloveso víceznačné: např., K rozlišení možných variant slouží příznaky *t* a *i* uvedené u jednotlivých heslových slov.

Podobně je zachycena i **reflexivita**, a to tak, že u heslového slova je podle potřeby uvedeno *se* nebo *si*, které pak slouží jako rozlišující příznak. Takto lze opět rozlišit případy jako . . .

První malá tabulka tab.0 tedy dává představu o vztazích mezi tranzitivou a intranzitivou a o četnostech sloves se *si* a *se* na základě sebraných cca 12 000 sloves.

Tabulka 0 – tranzitiva, intranzitiva reflexiva

```

-- i   : celkem 1700 sloves, tj.\,1700:119,42 = cca 15 \% \\
-- t   : celkem 6471 sloves, tj.\,6470:119,42 = cca 54 \% \\
-- se  : celkem 2780 sloves, tj.\,2780:119,42 = cca 24 \% \\
-- si  : celkem  572 sloves, tj.\,572:119,42  = cca  5 \% \\
-----\\
celkem          11523

```

Předběžná statistika valencí (a pádů)

V následujícím textu uvádíme v tab.1 předběžné údaje o četnostech jednotlivých přímých i předložkových pádů, jak se vyskytují u sloves v našem současném valenčním seznamu. Ve všech tabulkách jsou zatím jen absolutní četnosti, detailnější statistiky s ohledem na celkovou různorodost a komplikovanost valenčních vzorců budeme moci nabídnout až v dalším. Zatím nám počet různých valenčních vzorců v poměru k celému seznamu čítajícímu cca 12 000 sloves vychází na 4000, z nichž 2849 se vyskytuje s četností 1.

Tabulka 1 – souhrnné absolutní četnosti jednotlivých pádů

nominativ	11890
genitiv přímý	215
" předložkový	657
dativ přímý	295
" předložkový	193
akuzativ přímý	2341
" předložkový	589
lokál	1003
instrumentál přímý	878
" předložkový	392

Tab. 1 poskytuje základní a souhrnný přehled o distribuci přímých a předložkových pádů, které se vyskytují s jednotlivými slovesy samy o sobě, tj. je to základní přehled dvoumístných valencí tvořených na levé straně nominativem, který se implicitně objevuje téměř u všech sloves, a na pravé straně příslušným pádem z tabulky. Tabulka potvrzuje očekávanou převahu akuzativu a lokálu, následuje instrumentál a genitiv a jako poslední vychází dativ, u něhož, jak se dalo čekat, převažuje dativ přímý.

Tabulka 2 – Přehled výskytu konkrétních pádů

genitiv		dativ	
=====		=====	
čeho	161	komu	195

koho čeho	54	čemu	19
do čeho,	286	komu čemu	81
do koho čeho	38	k čemu	104
z čeho	222	ke komu čemu	56
z koho čeho	20	proti komu čemu	33
od čeho	13	-----	
od koho čeho	24	celkem	488

celkem	818		

akuzativ

=====

co	1461
koho co	880
na koho	57
na koho co	201
na co	217
o co	33
o koho co	24
pro koho co	28
za koho co	19
přes co	10

celkem	2930

lokál

=====

v čem	595
v kom čem	15
na čem	265
na kom čem	16
po čem	23
po kom čem	55
o čem	13
o kom čem	21

celkem	1003

instrumentál

=====

čím	842
kým čím	36
s kým	92
s čím	61
s kým čím	128
nad čím	18
nad kým čím	39
před kým čím	36
za kým čím	18

celkem	1256

větná doplnění

=====

\$(co, jak, že)	98
\$(že)	83
\$(aby)	28
koho \$(aby)	13
komu \$(že)	13

celkem	235

inf	77

celkem	77

Tab. 2 nabízí v porovnání s tab. 1 přehled četností konkrétních přímých a předložkových pádů získaných vytríděním z našeho seznamu. Lze z ní tedy vyčíst aspoň základní tendence u variant jednotlivých předložkových pádů a také rozdíly u přímých pádů. Např. u akuzativu (ale i u ostatních pádů) je vidět, že *co* je preferováno proti *koho—co*, což ukazuje na rozdílnou distribuci vzhledem k opozici životnost : neživotnost. Navíc jsou v tabulce uvedeny i základní údaje o infinitivní valenci a dále o větných doplněních a spojkách, které je uvozují.

Tabulka 3 – nejčtenější trojmístné valence

genitiv přímý - předložkový =====		akuzativ - genitiv př. =====	
koho ? do čeho	24	co ? z čeho	100
		co [z čeho]	57
dativ - genitiv předl. =====		co & z čeho	23
komu & do čeho	20	co ? do čeho	94
		co [do čeho]	39
dativ - akuzativ =====		co & do čeho	24
komu & co	322	koho co ? do čeho	52
komu & koho co	22	koho co ? z čeho	22
komu ? co	256	koho co [do čeho]	20
komu ? koho co	18	-----	
[komu] co	82	celkem	431
komu ? na co	19		
-----		akuzativ - dativ =====	
celkem	719	co ? k čemu	30
dativ - lokál =====		akuzativ přímý - předložkový =====	
komu ? v čem	30	co ? na co	57
		co [na co]	24
dativ - instrumentál =====		co ? na koho	32
komu ? čím	33	co & na koho	23
		co ? na koho co	22

		celkem	158
akuzativ - instrumentál =====		akuzativ - lokál =====	
co ? čím	250	co [v čem]	84
co [čím]	207	co ? v čem	36

koho ? čím	153	koho co [v čem]	53
koho [čím]	85	co ? (na čem, v čem)	22
koho co ? čím	264	-----	
koho co [čím]	256	celkem	195
co ? s kým	34		
co [s kým]	18		

celkem	1267		

instrumentál - lokál

=====

s kým ? o čem 26

Tabulka 3a – souhrn trojmístných valencí

genitiv přímý - gen.předložkový	24
dativ - genitiv	20
dativ - akuzativ	719
dativ - lokál	30
dativ - instrumentál	33
akuzativ - genitiv	431
akuzativ - dativ	30
akuzativ - lokál	195
akuzativ - instrumentál	1267
akuzativ přímý - ak.předložkový	158

Tab. 3 a 3a poskytují základní představu o nejčetnějších trojmístných valencích včetně údajů o konkrétních kombinacích přímých i předložkových pádů. K tomu poznamenáme, že pořadí, v němž jsou jednotlivé pády uváděny, je dáno zvoleným způsobem notace, takže si lze představit, že s ohledem na volný slovosled v češtině by obě uvedené tabulky mohly vypadat poněkud jinak, ale podstatu věci by to neovlivnilo. Pracujeme tedy se zápisem *věnovat komu & co*, ale stejně tak bychom mohli mít *věnovat co & komu*, informace o valenci a (v daném případě) o její obligatornosti tím není nijak dotčena. Tab. 3a ukazuje jasnou převahu dvojice *akuzativ-instrumentál* vyplývající z vyšší četnosti sloves s obecným významem *dělat něco něčím*. Druhé místo kombinace *dativ-akuzativ* rovněž není překvapující a je nepochybně dáno nemalým počtem sloves s valencí typu dávání či poskytování v širokém slova smyslu.

Z notace použité v tab. 3 lze také vyčíst rozdíly v distribuci obligatorních a fakultativních doplnění, např. zápis valence *komu & co* vyjadřuje, že jde o valenci obligatorní, zatímco zápis *komu ? co* pro nás znamená, že obě valence jsou v dané kombinaci fakultativní – mohou se spolu vyskytovat obě nebo jen jedna z nich. Zápis *[komu] co* pak

chápeme tak, že hranaté závorky vyznačují volné doplnění, které se u příslušného slovesa může a nemusí objevit a – v daném případě půjde s velkou pravděpodobností o volné dativy.

Tabulka 4 – počet významů u sloves

slovesa bez valence (rovnítka)	0:266
slovesa s jednou valencí (rovnítkem)	1:8429
se dvěma valencemi (rovnítky)	2:2196
se třemi	3:647
se čtyřmi	4:224
s pěti	5:73
s šesti	6:33
se sedmi	7:21
s osmi	8:6
s devíti	9:6
s desíti	10:3
s dvanácti	12:1
s patnácti	15:1
se sedmnácti: být	17:1
s padesáti osmi: jít (včetně frazémů)	58:1

0.5.10 Desambiguace – metody

1. techniky založené na pravidlech: DES a DES1
2. statistické techniky: bigramy, trigramy, Viterbiho algoritmus, probabilistické nekontextové gramatiky
3. využití neuronových sítí

0.6 Re prezentace významu

0.6.1 Lexikální význam – slova a slovní spojení

K významu obecně

Popis a definování významu představuje nejobtížnější oblast ZPJ. Přitom je zjevné, že bez vyřešení a zvládnutí této problematiky není možný další pokrok nejen v oblasti ZPJ, ale i v řadě oblastí AI – konkrétně se to týká reprezentace znalostí a inference.

Vezměme si např. sloveso *znamenat* – některá jeho užití se netýkají jazyka bezprostředně, tak např. ve větě

(v-v1) *Tyto stopy znamenají, že řidič začal brzdít pozdě.*

jde o to, že stopy na silnici jsou podkladem pro uvedený závěr. Na druhé straně ve větě (v-v2) *”Ploužit se” znamená jít pomalu.*

je tohoto slovesa použito způsobem, který nás bude dále zajímat. Konkrétně jde o definování (popis) významu slova pomocí jiných slov, tedy pomocí nějakého jazyka či přesněji řečeno metajazyka.

Obecně vzato, jaké máme možnosti, když se pokoušíme popisovat (definovat) význam slov nebo význam vět přirozeného jazyka? Bohužel to lze udělat jen tak, že k tomu použijeme zase jiného jazyka – metajazyka, jímž může být:

- týž nebo jiný přirozený jazyk
- nějaký formální jazyk, např. vhodný matematický nebo logický kalkul nebo jazyk sémantických rysů (sémů)
- z toho se vymyká ostenzivní způsob definování významu výrazů přír. jazyka: *to je auto, toto jsou klíče.* Přitom na ostenzivním definování významů je založeno učení se jazyku u člověka, mělo by tedy být přeneseno i do oblasti AI.

Další potíž spočívá v tom, že v přirozeném jazyce se běžně vyskytují věty jako: (v-v3) *Strážení poslanců ohrožuje demokracii.*, které ukazují, že jejich význam není **nezávislý na kontextu**. Otázka může být položena i jinak: lze popsat význam věty nezávisle na kontextu? Nebo má věta jen jeden význam, ale lze jí užít k různým účelům? Pokud by odpověď na první otázku byla kladná, mělo by to tu výhodu, že by bylo možno studovat význam věty detailně bez ohledu na veškeré komplikace spojené s jejich užíváním. Je-li tomu naopak a věty mají význam jen v závislosti na kontextu a **komunikační situaci**, pak zkoumání jazyka nelze oddělit od studia obecné lidské inference a situačního kontextu.

V dalším ukážeme, že **významy slov** lze zkoumat nezávisle na kontextu a že do jisté míry to platí i pro některé typy vět. Pokud nám půjde o kontextově nezávislý přístup, budeme mluvit o významu, jinak budeme pracovat s termínem **užití**. Zobrazení přiřazující větám PJ jejich sémantickou reprezentaci v podobě formulí PK1 nebo TIL budeme nazývat

– sémantickou interpretací,

zobrazení od sémantické reprezentace k finální reprezentaci znalostí (RZ) opět ve tvaru formulí PK1 nebo TIL předpokládá také zpracování **deiktických, indexických výrazů** a nazývá se

– **pragmatická interpretace**.

Je tedy rozdíl mezi normálními deskriptivními jmennými skupinami jako *chytrý student* a zájmennými skupinami typu *já, ty, on, my, tady, tam, ...* – u těchto jejich interpretace závisí na kontextu či komunikační situaci: ty určují, kdo je **mluvčí** a kdo **posluchač**. Jmenné skupiny s demonstrativy a posesivy typu *to auto, moje žena* či *ta moje žena* ovšem také závisí na kontextu, ale můžeme s nimi pracovat bez větších potíží stejně jako s deskriptivními jmennými skupinami (ev. je lze brát jako proměnné stejného typu).

Popis významů slov

Analogie se syntaxí – tam jsme zavedli slovní druhy a uvedli pravidla jejich kombinování do větších složek. Podobně to lze udělat se slovy a pokusit se je roztrždit významových tříd či sémantických polí, tj. nejprve si zavést vhodnou ontologii – tedy množinu tříd objektů, která představuje klasifikaci objektů universa U (podle Aristotela, 384-322 př.n.l.). Hlavní třídy objektů a jim odpovídajících jazykových výrazů podle Aristotela jsou:

- **substance**, tj. fyzické objekty
- **kvantity**, tj. např. čísla
- **vlastnosti**, tj. *červený, velký, krásný, milá*
- **relace** – typicky slovesa jako *milovat*, ale i *být otcem*
- **stavy**
- **události** – nejčastěji slovesa, stávají se, probíhají ve světě, poskytují strukturu pro interpretaci vět
- **akce** – to, co dělají agenti, činitelé, dá se na ně odkazovat zájmeny: (v-v4) *Zvedli jsme tu bednu. Byla to těžká práce.*
- **procesy**
- **situace** – jistý soubor okolností, situace v sobě subsumují události. Často jde o abstrakci úseku světa na určitém místě a v určitém čase: (v-v5) *Zuřili jsme a nadávali na fotbale,*
jak vidno, jde soubor akcí probíhajících na určitém místě a v určitém čase, např. fotbalový zápas. – **místo, locus** – *ve škole, tady, na rohu, doma*
- **pozice**
- **čas, tempus** – *teď, zítra, letos*
- **následek**
- **plány, záměry**

Naproti tomu ontologie, s níž se pracuje v PK1, zahrnuje jen individua a individuální proměnné, vlastnosti a relace – tedy entity prvního řádu.

Můžeme jít ještě dále a pokusit klasifikovat slova podle významu ještě detailněji

- dobře je to vidět na slovesech, u nich lze mít:
- slovesa pohybu: *jít, kráčet, utíkat, letět, vznášet se, ...*
- slovesa modální: *chtít, mít, moci, muset, smět, dát, ...*
- slovesa dicendi (sentiendi): *mluvit, říkat, říci, povídat, vědět*
- slovesa označující zpracování informace: *informovat, sdělovat, ...*
- slovesa označující emoce: *smát se, plakat, tesknit, ...*
- slovesa označující finanční transakce: *prodávat, kupovat, ...*

Levinové klasifikace sloves (Levin, 1995)

1. Slovesa tělesných funkcí a péče o tělo (275 syns.)
 - *potit se, trást se, omdlévat, bolet* - subjekt je neovládá, intransitivní.
 - *spát, chrápat, unavit se, mrznout*

- *mýt se, holit se, utírat se, oblékat se*
2. Slovesa změny (750 syns.), odpadkový koš, to, co nejde dobře jinam
 - *(z)měnit, modifikovat, upravit, adjustovat, lišit se*
 - *magnetizovat, elektrizovat, zvlhčit*
 - *zkrátit, prodloužit, zesílit, zeslabit, posílit, oslabit*
 3. Slovesa komunikace (710 syns.)
 - verbální: mluvit, koktat, blábolit
 - záměr mluvč.: prosit, žádat, nařizovat, děkovat, vyzývat, deklarovat
 - politika: vetovat, inaugurovat, omluvit
 - náboženské: kázat, modlit se
 - učit, přednášet, zkoušet, testovat
 - telefonovat, volat, faxovat, mailovat
 - zvířecí zvuky: řehtat, bučet, mňoukat, štěkat
 - hluky: skřípat, hrkat, vrzat, hučet, dunět
 4. Slovesa soutěžení (200 syns)
 - sporty: běžet, skákat, vrhat, házet, bruslit, lyžovat
 - hry: kopat, servírovat, útočit, vyhrávat, prohrávat, porazit
 - pískat, závodit, soutěžit
 5. Slovesa spotřeby, konzumace (130 syns)
 - požívání: jíst, pít, polykat
 - spotřeba: spotřebovávat, užívat, využít, použít
 6. Kontaktní slovesa (820 syns)
 - přidělat, připojit, přidat, přivázat, přivařit, při/upevnit
 - přikrýt, dotknout se,
 - oddělit, odříznout, odseknout
 - uchopit, stisknout, zmáčknout
 - pohladit, udeřit, praštit, trefit, zasáhnout
 - nést, strčit, manipulovat
 7. Kognitivní slovesa (? syns)
 - přemýšlet, uvažovat, usuzovat, pamatovat si, chápat, rozumět
 - dedukovat, inferovat, odhadovat, předpokládat
 8. Slovesa tvoření (250 syns)
 - mentálně: tvořit, vytvářet, vymýšlet, vynalézat,
 - umělecky: kreslit, malovat, rýt, tisknout
 - ze suroviny: péct, šít, vařit

9. Slovesa pohybu (500 syns)
 - na místě: hýbat se, otáčet se, kroutit se
 - v prostoru: pohybovat se, cestovat, běžet, utíkat, plazit se
 - v prostředí: plavat, létat
10. Slovesa emocí (?syns)
 - milovat, zbožňovat, nenávidět, bát se, postrádat, pohrdat
 - bavit, těšit, povzbuzovat, strašit, rozčilovat
 - tesknit, těšit se
 - cítit smutek, pociťovat radost
11. Statická (stavová) slovesa (200 syns), blízkost k adjektivům
 - být, mít: významy tohoto typu a podobné
12. Slovesa vnímání (percepce) (200 syns)
 - vidět, dívat se, hledět, zírat, slyšet, poslouchat
 - pozorovat, sledovat, hlídat
 - čichat, cítit, vonět, páchnout, smrdět
13. Slovesa vlastnění (300 syns)
 - mít, držet, vlastnit
 - dávat, dostávat, brát, vzít, získávat
 - dědit
 - krást, loupit
 - věnovat, darovat, poskytnout, uplácet, podplácet, korumpovat (?)
 - dodávat, odebírat, převádět
14. Slovesa sociálních interakcí (400 syns) zahrnují různé oblasti: právo, politika, ekonomika, rodina, náboženství, vzdělání
15. Slovesa počasí (66 syns)
 - pršet, lít, sněžit, padat (sníh), mžít, mrholit
 - blýskat se, hřmít
 - mračit se, zatahovat se, jasnit se

V průběhu SI vyvstává problém **víceznačnosti**:

– u slov, mají-li více významů než jeden. To zní jednoduše, ale jak zjistíme, že slovo má více významů? Můžeme se pokusit o svého druhu test: mějme slova *štěně, hlava, kulky, koule, kůň* a větu

(v-v6) *Já mám dvě koule a Honza má tři.*

Tuto větu lze jistě chápat dvěma způsoby, ale nikdy ne tak, že by v ní výraz *koule* označoval pokaždé něco jiného.

Na druhé straně slovo *kůň* ve větě

(v-v7) *Mám dva koně a Honza má tři.*

se nezdá být víceznačné, i když při každém jeho užití nemusíme být schopni rozlišit, zda se jím míní *klisna* nebo *hřibě*. To je jeden možný způsob, jak testovat naši intuici týkající se významů slov. U výrazu *koule* jde o víceznačnost, tedy přinejmenším o význam

K1 = geometrický objekt

a význam

K2 = varlata,

zatímco u *kůň* jde spíše o jistý druh *vágnosti*, kdy nemusí být jasné, zda máme na mysli *klisnu* nebo *hřibě*. Přesněji řečeno, platí mezi nimi a výrazem *kůň* významový vztah *hyponymie*. K němu se řadí další významové vztahy:

- *hyponymie* – *hyperonymie*
- *synonymie* – *antonymie*, např. *dobrý* : *zlý* apod.
- *meronymie* – *holonymie*, např. *nos* : *tvář* aj.

K tomu – viz **WordNet 1.5** a několik slov této organizaci slovníku a tomto typu slovníku obecně (instalace na FI, aisa, add module langtools, wn).

Podobný test lze navrhnout i pro slovesa, mějme větu:

(v-v8) *Měl jsem ji loni a Honza taky.*,

kde jistě můžeme rozlišit M1 = *vyspal jsem se tou slečnou* a proti tomu

M2 = *měl jsem chřipku*.

Je těžké si představit, že by tu mohlo o něco jiného než o plnou koordinaci. Proti tomu mějme:

(v-v9) *Políbil jsem Janu a Jirka taky.*,

i zde máme před sebou již zmíněnou *vágnost*, já jsem mohl *Janu* políbit na rty, kdežto *Jirka* jen na čelo. Místo, kam polibek přišel, není ve významu slovesa *políbit* explicitně specifikováno.

Souhrnně lze tedy říci, že v uvedených případech jde o *lexikální víceznačnost*, ale situace může být ještě komplikovanější, *víceznačnost* může mít *strukturní* povahu, může být způsobena syntaktickou strukturou věty:

(v-v10) *Kočky a fenky jsou spokojené a hrají si na zahradě.*

(v-v11) *Každý kluk má rád psa.*

(v-v12) *Mnoho lidí vidělo tu bouračku.* (10, 20, 50, 1000, ...)

Ve větě (v-v11) je jedna syntaktická struktura, ale SI můžeme mít více – a týkají se rozsahu kvantifikátorů (zkusit zapsat).

Ve větě (v-v12) jde o *vágnost* výrazu *mnoho* vzhledem k počtu lidí, kteří bouračku viděli. Zde můžeme mluvit o *sémantické víceznačnosti*.

Dále jsou tu případy jako:

(v-v13) *Já mám žízeň.*

(v-v14) *Ty se podíváš na to kolo.*

(v-v15) *Opravíme to tady.*

V nich je *víceznačnost* způsobena výrazy *já*, *ty*, *to*, *tady*, ..., kterým říkáme *deiktické* či

indexické. Jejich interpretace závisí na kontextu či na konkrétní komunikační situaci. Tento typ význačnosti můžeme charakterizovat jako význačnost **pragmatickou**. Lze pak uvažovat o **pragmatické funkci**, která vede od KS ke konkrétním hodnotám pro proměnné označené výrazy *já, ty, to, tady, ...* – jsou to patrně proměnné typu individuí (mluvčí, adresát, třetí osoba, ...).

0.6.2 Významy slov a slovníky

Významy slov a způsoby jejich popisu:

- pomocí synonym, např. v Oxfordském sl., SSJČ,
- pomocí definic, využití **generu proximum**, SSČ
- pomocí množiny vybraných primitivních výrazů daného přír. jazyka, např. *zabít* = *způsobit*, aby někdo zemřel – Hornby
- pomocí speciálního metajazyka: **sémantických rysů**, komponentová analýza – jednoduchý příklad:

muž = HUM, MASK, ADU

žena = HUM, FEM, ADU

chlapec = HUM, MASK, -ADU

dívka = HUM, FEM, -ADU

dítě = HUM, NEUT, -ADU

Další a podrobnější příklad – soubor možných rysů, příznaků, sémů (ČAJ):

T - tempus, čas, u substantiv jako "den, rok, leden, soumrak"

L - locus, místo, u substantiv jako "dům, chrám, světadíl, břeh"

BYT(ost) - např. "víla"

HUM(an) - člověk, např. "strejda, rada, bača", + M - muž, + F - žena

ANIM(al) - zvíře "pes, slon, velbloud"

PLANT - rostlina, např. "strom, kosatec"

QUA - vlastnost, např. "nespokojenec, povýšenec" + HUM

FEN(omén) - třeba "úkaz, zázrak"

ENT(ita) - "protiklad, argument"

OBJ(ekt) - předmět, např. "stůl, krb", ale také "dům", takže OBJ + L

INF(ormace) - např. "telefonát, článek, vzkaz, telegram"

EMOC(e) - třeba "cit, radost, strach, neklid, úsměv"

INS(trument) - nástroj, např. "nůž, šíp, hřeben"

MACH(ine) - stroj, aparát, zařízení, např. "počítač"

PROC(es) - např. "zážeh, postup, pokrok"

MOT(tion) - pohyb, např. "běh, let, pád"

AKT(ivita) - činnost, např. "boj, odboj, příchod"

MAT(eriál) - hlína, dřevo

B(ody) P(art), BP - prst, krk
ORG - organizace, instituce

Rysy lze kombinovat a jednomu výrazu jich přiřadit víc, viz třeba kombinaci "člověk" + "vlastnost", ev. i další. Lze zkusit i klasifikaci (hrubou) vlastností. Pokusme se zamyslet nad tím, že rysy mohou být hierarchické a že se díky tomu mohou dědit.

Typy slovníků:

- výkladové jednojazyčné, SSJČ, SSČ, Collins Cobuild, Webster, Oxford, jejich knižní a MRD verze.
 - vícejazyčné, překladové (Č-A, A-Č)
 - thesaury (Longman, WordNet 1.5, synonymické - SČS,
 - frazeologické, idiomů (SČFI)
 - jiné: dialektologické, etymologické, slangů, terminologické
- Ukázat aspoň ty hlavní.

0.6.3 Lexikální databáze

0.6.4 WordNet a sémantické sítě

Motivace

Standardním 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, sourozcům (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 fakturuální 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 se 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 uvedme 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, Fini-

tary 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 psycholinguista 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.

Struktura WordNetu

WordNet čili slovní síť je slovník podle autorů založený na psycholinguvistických 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, troponymie),
- 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).

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 vztahení 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*, jež lze charakterizovat jako vztah část – celek. Je v zásadě tranzitivní a antisymetrický a rovněž vede k budování hierarchických struktur.

Hyponymie/hyperonymie

Tyto vztahy uskupují substantiva tak, že tvoří lexikální dědičný systém. Popis významu substantivních synsetů (celkem asi 60 000) je ve WordNetu (obvykle) založen na nadřazeném výrazu (termu) doplněném o rozlišující příznaky (differentia specifica). 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 výtvar - 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 (výtvar, 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, happening (událost)	relation (vztah)
feeling, emotion (pocit, emoce)	shape (podoba, tvar)
food (potrava, jídlo)	state, condition (stav)

group, collection (skupina, soubor)
location, place (umístění, místo)
motive (motiv)

substance (substance, látka)
time (čas)

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

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.

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.

0.6.5 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, i když po lexikografické stránce vykazuje řadu chyb. V r.1997 se skupina lexikografů kolem P. Vossena z university v Amsterdamu rozhodla začít budovat síť slov pro tři vybrané západoevropské jazyky, a to v rámci projektu EuroWordNet-1, v jehož průběhu byla zároveň doplněna vrcholová ontologie a vytvořen soubor základních konceptů. Na

ten pak v r.1998 navázal EuroWordNet-2, do něhož byly zahrnuty další čtyři jazyky, z toho dva východoevropské.

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 adverbích 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 popisující 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 svému:

antonymům (dobrý : zlý)

hyperonymům (auto : dopravní prostředek)

hyponymům (pták : kanárek)

meronymům (dveře : zámek)

holonymům (ruka : tělo)

sourozencům (pes : vlk : kojot : hyena)

vyplývajícími 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:

a) 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 paradigmátům. Výchozí rámcovou představu o konstruktech ve vrcholové ontologii poskytuje Tab.1 výše.

b) 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. Před-

stavují 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.

c) 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.

d) 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.

EuroWordNet-2 – francouzština, němčina, čeština, estonština

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

a) 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 čítá 1000 synsetů, z toho je 700 substantivních a 300 verbálních.

b) Zachycení vnitřně jazykových vztahů (ILR) a vztahů ekvivalence v rámci základních konceptů (BC) pro němčinu, francouzštinu, estonštinu a češtinu. Výsledkem budou – de facto již jsou, – jádra wordnetů, každé v rozsahu 7500 synsetů, z toho je 5 000 substantivních a 2 500 slovesných synsetů. Adjektiva a adverbia zatím zůstávají stranou, ale s jejich zpracováním se počítá.

c) 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.

c) 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:

- 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 aktualizacího 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.

- 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 zachovávají vztahy specifické pro jednotlivé jazyky.

Obr.1 Architektura databáze EuroWordNet 2 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 (řídít) a díky tomu se nepřímou 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;

0.6.6 Budování české slovní sítě – českého WordNetu, dosa- vadní 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:

1. Výkladový slovník češtiny, což je pracovní název postupně budované lexikální data-

báze 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.

2.) Lingea Lexicon 2.0 (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. Tento zdroj 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.
3. 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:

- 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.
- Gramaticky i strukturálně značkový 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.
- extový korpus ESO budovaný na Fakultě informatiky v průběhu r. 1998 z novinových publicistických textů (1996-98), jeho aktuální rozsah činí 61 mil. českých slovních tvarů a jedna jeho verze je částečně lemmatizována.

0.6.7 Nástroje

Je zjevné, že popisovanou síť slov lze sotva budovat jen manuálně, má-li vzniknout v rozumném časovém úseku a s přijatelnými náklady. Při sestavování české sítě se tedy systematicky využívalo a využívá 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-1 a 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 dávkového zpracování dat českého wordnetu. V zásadě umožňují 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 ekvivaletů 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 zatřídování 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 souvškytu 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.

0.7 Sémantické reprezentace vět PJ

Zatímco pro popis syntaktické roviny existuje již v rámci počítačového zpracování přirozeného jazyka řada relativně propracovaných přístupů, jak jsme se snažili výše naznačit i pro češtinu, standardní techniky pro práci s významem vět a výpovědí prakticky neexistují. Následující úvahy budou proto mít poněkud volnější obrysy a půjde v nich spíše o mapování některých aktuálních směrů výzkumu.

Povšimneme si sémantických reprezentací, otázek reference a aplikace principu kompozicionality. Budeme věnovat pozornost algoritmu překladu syntaktických reprezentací na sémantické a případně i některým otázkám spojeným s víceznačností.

Máme-li vysvětlit schopnost uživatele jazyka rozumět výrazům přirozeného jazyka, musíme postulovat existenci nějaké vnitřní reprezentace významu výrazů přirozeného jazyka. I když v současnosti nelze dost dobře odpovědět na otázku, jakou konkrétní podobu mají u člověka tyto vnitřní reprezentace významu, z povahy jazykové komunikace a na základě introspekce lze dospět k závěru, že bez postulování sémantických reprezentací se neobejdeme.

Mají-li SR splňovat svůj účel, měly by vyhovovat aspoň následujícím požadavkům:

1. SR by měly umožňovat **jednoznačné** zachycení významů výrazů přirozeného jazyka (dále PJ),
2. SR by měly umožňovat postížení **synonymie** (parafráze) výrazů jazyka, tj. situace, kdy různým větám odpovídá jeden význam – jedna SR. Máme tu na mysli např. situace, kdy následující otázky lze zodpovědět jedním způsobem:
 - (a) *Kdo měl poměr s ředitelovou ženou?*
 - (b) *Kdo spal s ženou ředitele?*
 - (c) *Byl to údržbář.*
3. SR by též měly umožňovat přirozené postížení **homonymie** jazykových výrazů, tj. situaci, kdy jedné větě odpovídá více významů a tudíž jí bude přiřazeno více SR.

Při zkoumání vztahů mezi výrazy jazyka a jejich odpovídajícími SR lze postupovat ve dvou směrech:

1. od výrazů jazyka k hledaným odpovídajícím SR – tento přístup můžeme charakterizovat jako **analýzu**,
2. od SR (za předpokladu, že existují indukční pravidla jejich formování) k výrazům jazyka – tento přístup charakterizovat jako **syntézu**.

V následujících úvahách se budeme zaměřovat spíše na **syntézu**, ačkoli na této úrovni výkladu není uvedená distinkce podstatná. Svého plného významu nabývá až v okamžiku, kdy se začneme zabývat implementovatelnými algoritmy.

Pokusíme se tedy vést paralelu mezi postulovanou uživatelskou vnitřní reprezentací významu výrazů přirozeného jazyka a tím, co budeme dále nazývat *sémantickou reprezentací* výrazů přirozeného jazyka. Zde bude klíčovou otázkou, jakých prostředků k budování sémantických reprezentací (dále SR) použijeme.

0.7.1 Formální aparát pro SR – charakteristika TIL

V současných lingvistických teoriích se významy výrazů (slov, slovních spojení, vět) přirozeného jazyka nejčastěji popisují na základě aparátu predikátové logiky 1. řádu (Winoograd, 1972, Gazdar, Mellish, 1989). Podle našeho názoru lze však pokládat za **dostatečně vyjasněné** (viz např. Tichý, 1976, Svoboda, Materna, Pala, 1979, Materna, Pala, Zlatuška, 1989), že predikátová logika 1. řádu (dále PL1) není neadekvátnějším nástrojem pro zachycení SR, neboť se jí nedostává potřebné vyjadřovací síly – řadu významů běžně vyjadřovaných v kterémkoli přirozeném jazyce nelze prostředky predikátové logiky 1. řádu dostatečně systematicky zachytit. Citované práce přesvědčivě argumentují, že vhodnější k těmto účelům a empiricky adekvátnější je aparát intenzionální logiky, který ve variantě, jíž budeme dále věnovat pozornost, bývá charakterizován jako tzv. **transparentní intenzionální logika** (dále TIL, Tichý, 1976, Tichý, 1988, Materna, Pala, Zlatuška, 1989²).

a) TIL je logický systém založený na určité modifikaci (viz zejména dále pod b)) **typovaného lambda kalkulu**. Lambda kalkul je logický aparát, který umožňuje manipulaci s funkcemi. Rozumná interpretace tohoto aparátu, který má obecně velké uplatnění v matematice a informatice, je umožněna principem teorie typů, který tvorbu funkcí omezuje na základě výstavby tzv. hierarchie typů a podle něhož funkce nemůže být aplikována např. na sebe samu. Typovaný lambda kalkul manipuluje s funkcemi v souladu s principem teorie typů. Tím, že je založen na neomezené hierarchii typů, je typovaný lambda kalkul vhodným aparátem k překonání nedostatečné expresivity, jaká je vlastní např. PL1.

I jiné systémy než TIL, zejména jiné intenzionální logiky, jsou založeny na aparátu typovaného lambda kalkulu. Pokud však modifikují tento aparát, pak nikdy ve smyslu b), resp. c) (viz dále).

²V následujícím výkladu se budeme opírat o řadu formulací z této práce. Podrobnější charakteristika formálního aparátu TIL je uvedena v příloze v odd. 0.9.5

b) TIL je transparentní systém, tj. pro TIL není formální aparát reprezentující způsoby, jakými jsou konstruovány objekty, předmětem studia, nýbrž pouze prostředkem ke studiu těchto konstrukcí.

Tímto rysem se TIL odlišuje od všech soudobých logických systémů: zatímco v TIL je formální výraz označením konstrukce, je pro stoupence formalismu tento výraz bezprostředním jménem konstruovaného objektu. Na triviálním příkladu lze tento rozdíl ukázat takto:

	formální pojetí	TIL
výraz	3 + 5	3 + 5
sémantika složek výrazu	číslo 3 číslo 5 operace sčítání	číslo 3 číslo 5 operace sčítání
sémantika výrazu	číslo 8	konstrukce, tj. určitý způsob, jakým uvedené složky spolupracují na vytvoření objektu

Vidíme, že pro formalistu neexistuje sémantický *mezistupeň* mezi objekty označenými složkami složeného výrazu a objektem *výsledným*. Pro TIL je sémantika výrazu dána tím, že způsob, jakým je tento výraz strukturován, zobrazuje strukturu konstrukce, jejímiž složkami nejsou složky jazykového výrazu, nýbrž objekty těmito složkami označené. Jak ukázal autor TIL v řadě statí (a zejména ve své monografii, Tichý, 1990), vede ignorování pojmu konstrukce k řadě chyb, nedorozumění i pseudoproblémů.

c) TIL nepreferuje jistá *vybraná slova* jako tzv. *logická slova*, jež by údajně určovala charakter logiky.

Také tento rys je specifický pouze pro TIL (souvisí s rysem b)). V ostatních, formálně budovaných systémech se vždy setkáváme s množinou vyčleněných *konstantních* výrazů, které jsou *logické* a které jediné zajišťují odlišení logicky pravdivých vět, logického vyplývání, logické ekvivalence od ostatních (zřejmě na empirii závislých) vlastností a vztahů. Tak ve výrokové logice jsou logickými slovy logické (výrokové) spojky, v PL1 k nim přistupují kvantifikátory, resp. identita. Tato *logická slova* jsou navíc chápána jako tzv. *nevlastní symboly*, tj. interpretaci jim není přiřazován soběstačný význam; význam je přiřazován jen celým složeným výrazům, které je obsahují.

Z tohoto hlediska např. věta

(15) *Pavel je starší než Petr.*

není logicky ekvivalentní větě

(16) *Petr je mladší než Pavel.*,

protože analýza těchto vět v PL1 dává

(15') $\text{St}(\text{Pavel}, \text{Petr})$, resp.

(16') $\text{Ml}(\text{Petr}, \text{Pavel})$,

takže se nemůžeme opřít o žádné *logické slovo*, na jehož základě bychom mohli odvodit ekvivalenci (15) a (16). Samozřejmě, i PL1 odhalí logickou souvislost těchto vět tím, že zavede *významový postulát*

(17) $\forall xy (\text{St}(x,y) \equiv \text{Ml}(y,x))$

a prohlásí, že (15') je ekvivalentní s (16') za předpokladu (17). Ale (17) je z hlediska intuice logicky pravdivá věta, takže ji nepokládáme za zvláštní předpoklad. Jenže (17) nemůže být z hlediska PL1 logicky pravdivá věta: aby jí byla, musela by být pravdivá ve všech strukturách. Snadno však najdeme takovou strukturu, v níž (17) neplatí; stačí za U zvolit např. množinu přirozených čísel a za relace, jež budou interpretací přiřazeny St, resp. Ml, relace $>$, resp. \geq .

Další charakteristiky TIL se týkají aplikace TIL na analýzu přirozeného jazyka.

d) TIL aplikována na analýzu přirozeného jazyka se stává sémantikou založenou na pojmu *možných světů* (*possible worlds semantics*).

Tento rys sdílí TIL s nejrozšířenějšími aplikacemi logických systémů na analýzu přirozeného jazyka. Myšlenka využít možných stavů světa, popř. časových okamžiků k definování intenzí jako logicky manipulovatelných objektů se stala v soudobé logické sémantice převládající ideou.

Poznámka:

Termín *možný svět* byl převzat z Leibnize a poprvé v zárodečné moderní podobě použit R. Carnapem. Někdy se mluví i o *množině indexů* (Montague aj.), do níž jsou vedle možných světů a časových okamžiků zařazovány některé další parametry (ponejvíce pragmatické povahy). S kategorií možných světů pracuje i tzv. finská logická škola (J. Hintikka aj.).

e) Univerzum je v TIL chápáno jako množina společná všem možným světům.

Tento rys je charakteristický zejména pro TIL; ve většině ostatních koncepcí se uvažuje vedle možných světů i o *možných individuích*, tj. populace individuí je obecně různá v různých možných světech. Tento zdánlivě samozřejmý předpoklad (v některém možném světě existuje Pegas, v jiném ne) byl koncepcí TIL přesvědčivě vyvrácen.

f) Fregeho (Churchovo) rozlišení vztahu denotace jakožto označování (reference) a vztahu vyjadřování *smyslu* je v TIL zrušeno a nahrazeno jiným schématem.

Také tento rys nalezneme u malého počtu jiných systémů; většinou je denotace (označení, pojmenování, reference) vztažena k extenzím a intenze jsou chápány jako výsledek způsobu vyjádření.

Vedle těchto rysů charakteristických pro TIL je třeba se zmínit o specifickém deduktivním aparátu, který je obdobou syntaktického důkazového aparátu v PL1, ale je přizpůsoben transparentní koncepci; neklade důraz na *axiómy*, je generalizací Gentzenovy *přirozené dedukce* (s touto teorií se lze seznámit např. v Janákové práci, (1973)) na teorii typů a je velmi účinný. Nejjednodušší aplikace tohoto aparátu byla u nás realizována v systému ADAM pro reprezentaci znalostí na počítači CYBER 172. (Viz T. Chrz, 1984).

0.7.2 Formální aparát – TIL a teorie typů

Předchozí úvahy nás vedou k hledání formálního aparátu vhodného pro sémantickou analýzu výrazů PJ. Jak jsme už naznačili, za takový nástroj pokládáme zmíněný již TIL. Základními rysy systému TIL jsou:

1. schopnost systematicky překračovat omezení platná v predikátové logice 1. řádu (extenzionální sémantice);
2. důsledný intenzionalismus a z něho vyplývající schopnost přesného definování intenzí a zacházení s nimi;
3. vzhledem k přirozenému jazyku disponuje TIL větší expresivní silou – což plyne z bodu 1.

Podrobnější charakteristiku systému TIL a jeho vlastností, díky nimž je tak zajímavý a vhodný pro sémantickou analýzu PJ, uvádíme samostatně v příloze **Teorie typů**. I zde primárně vycházíme z citované již práce Materna, Pala, Zlatuška, 1989.

0.7.3 Sémantická analýza výrazů PJ

Jedním z hlavních cílů sémantické analýzy PJ je ukázat, jak význam složeného výrazu může být odvozen z významů jeho složek. Je patrné, že velmi vhodným nástrojem k tomu jsou konstrukce uvedené výše.

Analyzovat sémanticky výraz přirozeného jazyka (větu) znamená **nalézt** konstrukci, která je tímto výrazem vyjadřována. Tuto konstrukci můžeme pak pokládat za **sémantickou reprezentaci** analyzovaného výrazu. Pokud však výsledkem analýzy není jednoznačná konstrukce, vzniká potřeba konstrukci **standardizovat**, což se neobejde bez zavedení tzv. „linguistic constructions“ (Hajičová, Materna, Sgall, 1988).

Zajímá-li nás přirozený jazyk jako např. čeština a je-li dána epistémická báze B_L příslušející k tomuto jazyku, lze při budování konstrukcí vyjadřovaných větami tohoto jazyka – budeme jej značit L – postupovat zhruba následovně:

1. Mějme následující českou větu:
(v18) *Studentka Alena si myslí, že ministr financí je hezčí než ministr zahraničí.*
2. Nejprve se pokusíme zjistit, která slova z (v18) označují atomy nad B_L . Můžeme to učinit tak, že nahlédneme do sémantického slovníku, v němž pro jednoduchost najdeme u příslušných slovních tvarů jejich odpovídající typové charakteristiky.
 - Musíme však počítat s tím, že některá slova v L mohou patřit současně do více kategorií, to platí např. o slovese *být* a dalších. Je potřeba přihlédnout i k okolnosti, že i některé gramatické kategorie (rysy) mohou označovat atomy nad B_L – gramatické časy, vidy, gramatické číslo.
 - To, co následuje, lze pokládat za minimální fragment takového slovníku. Samostatným problémem je stavba takového slovníku a způsob jeho vytváření – jeden pokus týkající se českých sloves lze nalézt v práci B. Podlezlové-Koželouhové (1974). Další velmi zajímavou analýzu týkající se českých sloves a slovesného času předložila J. Koukolíková (1988).
3. Víceslovné výrazy pokládáme pro jednoduchost za celky.
studentka Alena: \mathbf{A}/ι – nálepka individua
myslet si: $\mathbf{M}/(oio_{\tau\omega})_{\tau\omega}$ – vztah mezi individuem a propozicí
ministr financí: $\mathbf{F}/\iota_{\tau\omega}$ – individuální koncept
hezčí než: $\mathbf{Hn}/(ou)_{\tau\omega}$ – vztah mezi dvěma individui
ministr zahraničí: $\mathbf{Z}/\iota_{\tau\omega}$ – individuální koncept.
4. Další krok spočívá v nalezení konstrukce vyjadřované větou (v18) a tabulky funkce, jež je touto konstrukcí konstruována. Protože (v18) je souvětí, začneme nejprve analyzovat vedlejší větu, která je uvozena spojkou *že*. \mathbf{Hn} je vztah mezi individui, \mathbf{F} a \mathbf{Z} však nejsou individua. Budou-li ale aplikována na nějaký svět W v okamžiku S , mohou vytvořit ι -konstrukce, tj. hodnotou \mathbf{F} ve světě W a okamžiku S může být ta určitá osoba, např. právě *Václav K.* a podobně hodnotou \mathbf{Z} může být třeba *Jiří D.* Aplikace \mathbf{F} a \mathbf{Z} na svět W v okamžiku S se uskuteční prostřednictvím ω -proměnné w (možných světů) a τ -proměnné t časových okamžiků. Podobně postupujeme u atomu \mathbf{Hn} , což vede ke konstrukci:
 (K1) $(\mathbf{Hn}_{wt}(\mathbf{F}_{wt}, \mathbf{Z}_{wt}))$.
 Jak si lze bez větších obtíží ověřit, výsledná o -konstrukce není uzavřená, obsahuje výskyty volných proměnných w a t . Tato konstrukce v -konstruuje pravdivostní hodnotu v závislosti na možném světě W a okamžiku S . Další krok spočívá nyní v tom, že použitím λ -operátoru se zbavíme volných výskytů proměnných w a t , a tak dostaneme konstrukci (K2), která již konstruuje propozici:
 (K2) $\lambda w \lambda t (\mathbf{Hn}_{wt} \mathbf{F}_{wt} \mathbf{Z}_{wt})$.
 Přidání atomů \mathbf{M} a \mathbf{A} vede již ke konstrukci (K3), která je vyjadřována naší větou (v18).
 (K3) $\lambda w \lambda t (\mathbf{M}_{wt} (\mathbf{A} (\lambda w \lambda t (\mathbf{Hn}_{wt} \mathbf{F}_{wt} \mathbf{Z}_{wt}))))$.

Vidíme, že (K3) konstruuje objekt $o_{\tau\omega}$ – tedy propozici, což je funkce, která každému možnému světu W v okamžiku S přiřadí nejvýše jednu pravdivostní hodnotu. V těch možných světech a těch okamžicích, v nichž si studentka Alena myslí, že platí propozice konstruovaná konstrukcí (K2), je přiřazenou hodnotou \mathbf{P} , v ostatních světech a okamžicích je touto hodnotou \mathbf{N} . Konstrukce (K2) konstruuje propozici, v jejíž pravdivost studentka Alena věří a která nabývá hodnoty \mathbf{N} v těch světech a okamžicích, v nichž individuum, které je ministrem financí (\mathbf{F}_{wt}), a individuum, které je ministrem zahraničí (\mathbf{Z}_{wt}), jsou v relaci, jež je hodnotou vztahu \mathbf{Hn} . V těch světech a těch okamžicích, v nichž zmíněná individua v této relaci nejsou, nabývá propozice hodnoty \mathbf{N} . Posléze v těch světech a těch okamžicích, ve kterých žádné individuum není ministrem financí nebo ministrem zahraničí (nebo obojí), je propozice nedefinována. Podotkněme k tomu, že v aktuálním světě je tato propozice v přítomnosti definována: české větě vyjadřující konstrukci (K2) lze přiřadit pravdivostní hodnotu. Dodejme ještě, že pravdivost propozice konstruované (K3) nezávisí na pravdivosti propozici konstruované (K2).

0.7.4 Nástin algoritmu sémantické analýzy

Nyní nás budou zajímat možnosti algoritmizace sémantické analýzy výrazů PJ popsané výše, a to s cílem dospět k sémantickému analyzátoru, který by v úzké návaznosti na již popsaný syntaktický analyzátor budoval pro vstupní české věty jejich odpovídající SR. Navazujeme tu na dřívější experimentální syntakticko-sémantický analyzátor pro omezenou podmnožinu českých vět, který byl napsán v programovacím jazyce LISP (Pala, Materna, 1976, Palová-Vaníčková, 1978, Čihánek, 1978, nejnověji se o implementaci jednoduchého sémantického analyzátoru v PROLOGU pokusila Koukolíková, 1988).

Ať už zvolíme přístup *rule-to-rule* (každému syntaktickému pravidlu je přiřazeno odpovídající pravidlo sémantické) či postup *sekvenční*, kdy se nejprve provádí syntaktická analýza, jejímž výsledkem je stromový graf reprezentující syntaktickou strukturu vstupní věty, v každém případě musíme počítat se dvěma okruhy vstupních dat:

1. s informacemi o syntaktické struktuře vstupní české věty v podobě vhodného stromového grafu, který např. může být výstupem z výše popsaného syntaktického analyzátoru. U přístupu *rule-to-rule* by šlo o tytéž informace, z technického hlediska by se s nimi ovšem zacházelo poněkud jinak, neboť některé kroky by se prováděly prakticky současně;
2. s vhodnou formou sémantického slovníku, který v zásadě může obsahovat do značné míry stejné lexikální jednotky jako slovník syntaktický, ovšem s poněkud jinými údaji. Lze ovšem mít i slovník jeden, který při vhodném uspořádání může sloužit oběma částem analýzy, ale to je otázka do značné míry technická a implementační, kterou se zde nebudeme podrobněji zabývat. Zde budeme vycházet z toho, že lexikálním jednotkám jsou v sémantickém slovníku přiřazeny vhodné **typové popisy**

a že tam jsou i další potřebné údaje týkající se např. kvantifikátorů, logických spojek, předložek, částic ap.

Vlastní sémantická analýza může začínat testováním uzlů syntaktického stromu a rysů v seznamech připojených k uzlům. Jak uzly tak rysy obsahují údaje předurčující celkový průběh sémantické analýzy, je v nich totiž obsažena informace, že např. věta je tázací, je v ní budoucí čas, hlavní sloveso je negováno apod.

Po provedení testů tohoto druhu lze standardním způsobem založit kořen sémantického stromu ($o\omega$) a jeho obligatorní potomky $\lambda\omega$ a o . Od časových okamžiků τ budeme zatím odhlížet, i když jejich začlenění není spojeno s žádnými zvláštními komplikacemi.

Dalším význačným krokem je analýza slovesné skupiny ve větě. Začíná tím, že pro hlavní sloveso analyzované věty se v sémantickém slovníku najde jeho typová charakteristika. Poté je průběh analýzy do značné míry závislý na výsledcích syntaktické analýzy: dostaneme-li např. ze syntaktické analýzy údaje o tom, že slovesná skupina v analyzované větě je tvořena sponovým slovesem *být* a jmennou skupinou v nominativu, je slovesné skupině bez dalšího testování přiřazen typ vlastnosti, tj. $(o\alpha)_{\tau\omega}$ (kde α je jakýkoli typ).

Poté se hledají adverbia míry a způsobu. Jsou-li nalezena, připojí se pod slovesnou skupinu a s použitím operace aplikace se vytvoří celkový typ slovesné skupiny. Pokud jde o typy těchto adverbí, lze pro začátek vyjít z toho, že označují objekty, které mohou být charakterizovány jako vlastnosti vlastností, tj. mohou být spojovány s podobnými objekty jako níže zmíněná adjektiva. Je-li ve větě nalezena (při syntaktické analýze) adverbíální skupina s rysem místa nebo času, založí se pro ni v sémantickém stromu samostatný uzel. U adverbí času jako *včera*, *dnes*, ... lze počítat s tím, že označují $(o\tau)\tau$ -objekty, i když v této souvislosti se nevyhneme podrobné sémantické analýze gramatických časů a vidů u značného počtu českých sloves, jak je naznačena u Tichého (1980).

Následuje v podstatě nejobtížnější fáze analýzy, jíž je analýza jmenných skupin ve větě. Nejprve se testuje, zda počet argumentů indikovaných typem slovesné skupiny se shoduje s počtem jmenných skupin v syntaktickém stromu věty. Je-li výsledek testu negativní, analýza se vrací zpět ke slovesné skupině, u níž se v sémantickém slovníku snažíme najít další typ. Pokud uspějeme, výše popsaný proces se opakuje. Je-li výsledek testu na počet jmenných skupin ve větě pozitivní, přistoupí se již k analýze jmenných skupin, která u každé jednotlivé skupiny probíhá nejprve zdola nahoru, tj. ve slovníku se vyhledají typy složek tvořících jmennou skupinu (např. *A N – chytrý poslanec*).

Nejprve je potřeba vyrovnat se s typy adjektiv. Obecně lze počítat s tím, že adjektiva označují $((o\alpha)(o\alpha)_{\tau\omega})_{\tau\omega}$ -objekty pro nějaký typ α : jsou to tedy funkce, které každému stavu světa přiřadí funkci, jež každé vlastnosti α -objektů přiřadí určitou třídu α -objektů, což se uplatní při analýze výrazů jako

(v19) *Můj kamarád je chytrý poslanec.*

Druhou možností je, že adjektiva označují $(\iota(o\iota)_{\tau\omega})_{\tau\omega}$ -objekty, což se vztahuje např. k vý-

razům *nejdemokratičtější prezident* nebo *ten chytrý poslanec*.

Poznamenejme však, že u výrazů (jmenných skupin) obsahujících demonstrativa či posesiva se nabízí možnost typovou analýzu minimalizovat, neboť tato zájmena vcelku spolehlivě signalizují, že jmenné skupiny, které je obsahují, lze bezpečně analyzovat jako výrazy označující individuální objekty.

Pak se postupem shora dolů (počínaje uzlem NP) činí pokus sestavit výsledný typ celé jmenné skupiny, jenž byl již predikován typovou charakteristikou slovesa získanou v předchozím průběhu analýzy.

Je-li výsledek analýzy všech příslušných jmenných skupin ve větě pozitivní, je sestaven sémantický strom analyzované věty spolu s jeho *linearizací*, která je hledanou konstrukcí, již analyzovaná věta vyjadřuje. Tuto konstrukci pak můžeme pokládat za *sémantickou reprezentaci* analyzované vstupní věty.

Po takto provedené analýze mohou nastat dvě situace:

1. Získaná konstrukce (SR) neobsahuje žádné volné proměnné a je tudíž **uzavřená**. V tom případě lze celou analýzu pokládat za definitivně a úspěšně ukončenou.
2. Výsledná konstrukce (SR) obsahuje volné proměnné a je tedy **otevřená**. Nastane-li tento případ, je nutno přejít k analýze pragmatické, která by měla poskytnout chybějící údaje potřebné k získání uzavřené konstrukce (SR) (viz dále).

Jak lze vidět z předchozího, naznačený algoritmus se přirozeně člení do čtyř modulů, které byly v Čihánkově programu (Čihánek, 1978) realizovány jako lisovské funkce:

1. **přípravný modul I** – v něm se analyzují slovesné časy, větná negace (spojená s finitním slovesným tvarem) a provádějí se přípravné akce pro analýzu slovesné skupiny;
2. **slovesný modul** – analyzuje slovesnou skupinu věty a adverbia patřící k hlavnímu (finitnímu) slovesu ve větě, též výrazy s významem místa a času a případně i další;
3. **přípravný modul II** – provádí přípravné akce pro analýzu jmenných skupin (nastavení hodnot programových proměnných potřebných pro koordinaci činnosti slovesného a jmenného modulu);
4. **jmenný modul** – provádí sémantickou analýzu jmenných skupin ve vstupní větě, tj. sestavuje na základě syntaktických informací jejich výsledné typy a začleňuje je do typu získaného již dříve při analýze slovesné skupiny věty. Dokončuje celou analýzu, tj. vytváří výsledný sémantický strom a jemu odpovídající linearizaci hledané konstrukce – sémantické reprezentace vstupní věty a podle potřeby i jejich grafické podoby.

0.7.5 Poznámky k sémantické roli jmenných skupin

Typickou funkcí singulární **np** ve větě zhruba je označovat nějaký objekt univerza promluvy, který je relevantní v dané komunikační situaci. Z hlediska počítačové analýzy je problémem skutečnost, že **np** může být ve větě **víceznačná** nebo **neurčená**. Nicméně lze počítat s jistými základními regularitami, pokud jde o hlavní funkce **np**. (V těchto úvahách počítáme spíše s extenzionálním pojetím reprezentace objektů v reprezentaci daného výseku světa, i proto, že konkrétní počítačové reprezentace výseků světa zatím plně intenzionální pojetí neumožňují, neboť nejsou vybaveny koncepty (intenzemi) jako rozpoznávacími procedurami.)

1. **neurčité np**, např. *nové kolo, nějaké děti, tři docenti*, jsou obvykle extenzionálně chápány tak, že označují specifický objekt nebo jejich množinu, u nichž se předpokládá, že jsou pro adresáta nové. Z hlediska algoritmického popisu (a odpovídajícího počítačového programu) to znamená, že v dané reprezentaci světa se vytvoří nový vnitřní symbol, který bude označovat příslušný objekt, a přidá se do aktuální reprezentace daného výseku světa. Máme-li větu
(v20) *Karel si koupil nové kolo.*,
do aktuální reprezentace světa se poznamená něco jako
kolo(k1)
nový(k1)
vlastnit(karel, k1).
2. dále se neurčitých **np** užívá v nespécifických kontextech pro označení objektů, které mohou nebo nemusí existovat, např. ve větě
(v21) *Karel si chce koupit nové Shimano.*
jde podle extenzionalistů o tzv. **nepřímý** (opaque, oblique) kontext, který je spojen se slovesy jako *věřit, chtít, myslet, doufat, přát si* aj.
3. neurčitých **np** lze též (extenzionálně vzato) užít genericky k označení třídy objektů jako např. ve větě
(v22) *Nové kolo vyžaduje pravidelnou údržbu.*
– Typické jsou konstrukce se slovesem *být* nebo *stát se*, jichž se často užívá k vyjádření skutečnosti, že (extenzionálně) daný individuální objekt patří do nějaké třídy (má nějakou vlastnost), např.
(v23) *Ta hromádka zkrouceného kovu je nové kolo.*
(v24) *Shimano XJ je nové kolo.*
4. určité **np** mají někdy užití, které je dosti podobné neurčitým **np**, např. **np** ve (v25) označuje konkrétní objekt, ve (v26) jde o užití generické
(v25) *Karlovi se přestalo líbit to nové Shimano, co si koupil.*
(v26) *Jaguár je příbuzným leoparda, který žije v Jižní Americe.*

5. Určité **np** nezřídka hrají roli deskripcí, které v terminologii intenzionální sémantiky označují individuální koncepty (tzv. offices), např.
(v27) *Výrobce tohoto kola by měl být volán k odpovědnosti.*
Np tohoto typu obvykle umožňují identifikovat denotát bez větších komplikací, pro extenzionální pojetí však představují nemalé potíže.

6. zájmena, resp. koncovky verba finita, odkazují zpět k individuálním objektům které byly v promluvě uvedeny předchozími **np**. Tak např.
(v28) *Karel si v dražbě koupil staré Shimano XJ.*
Bylo už pěkně ojeté.
– O poněkud jinou situaci jde v následujícím případě, i když se tváří do jisté míry podobně jako předchozí
(v29) *Karel si chtěl opatřit láhev s džinem.*
Doufalø, že mu bude uklízet byt.
Zde se zájmenná a „koncovková“ reference týká individuálního konceptu a individua, navíc **np láhev s džinem** je víceznačná, takže zjištění korektní reference prostřednictvím koncovky 3. os. sg. může být velmi nesnadné, jestliže daná reprezentace světa neobsahuje žádné údaje o pohádkových bytostech.

7. poznamenejme však, že **np** se objevují též v konstrukcích typu
(v30) *Žádný řidič nepřipustí, že je horší než nějaká ženská.,*
v nichž ovšem nelze mluvit o referenci jako takové – tyto **np** vyžadují jiný typ analýzy, neboť se vztahují k logickým kvantifikátorům (obecnému a existenčnímu) a navíc jsou ještě spojeny s operátorem negace.

8. V dosud uvedených příkladech jsme věnovali pozornost výlučně oznamovacím větám. U otázek a rozkazů lze očekávat interpretaci neurčitých **np** jako deskripcí objektů, které by adresát měl identifikovat v průběhu procesů tázání se a odpovídání a rozkazování a provádění rozkazů, např.
(v31) *Je ta tvoje kniha v pokoji na stole?*
(v32) *Dej tu jeho knihu do pokoje na stůl!*

Pokusili jsme se naznačit některé základní funkce **np** v oblasti reference a nyní vzniká otázka, jak se s těmito otázkami vyrovnat v rámci počítačového modelu porozumění PJ. Situace je o to nepříjemnější, že syntaktické prostředky neposkytují příliš často jasná vodítka pro rozpoznání příslušné funkce **np** (jako je tomu např. u konstrukcí se slovesem *být*. Adresát je většinou odkázán na znalost tématu konverzace a kontextu a z nich musí odvodit příslušnou funkci **np**. V počítačových modelech se však zjednodušeně počítá jen s **np** a zájmeny, které se vyznačují konkrétní referencí, dále s tím, že neurčité **np** jsou specifické nebo v případě otázek a rozkazů nespecifické. Přes tyto simplifikace jsou problémy s interpretací **np** v netriviálních kontextech značné.

0.7.6 Referenční role funkční perspektivy větné

Obvyklým cílem pronesení oznamovací věty je sdělit novou informaci, která adresátovi není dosud známa. Aby si adresát mohl integrovat tuto informaci do své zásoby existujících znalostí, může mu mluvčí poskytnout jisté množství známé informace, kterou již adresát disponuje (Sgall, Hajičová, 1985). Např. ve větě (v33) *Posledně jsem mu vysvětloval principy českého slovosledu.* adresát pravděpodobně ví, že výraz *mu* odkazuje k jednomu z mých studentů a že jsem to byl *já*, kdo vysvětloval. Nová informace se pak týká toho, co bylo vysvětlováno.

V jakém smyslu je distinkce nového (rématu) a známého (tématu) (Firbas, 1971, Sgall, Hajičová, Buráňová, 1980) relevantní pro (sémantickou) interpretaci jmenných skupin? Za předpokladu, že vedeme konverzaci s partnerem, který je kooperativní, může adresát očekávat, že nová a známá informace bude nějak vyznačena a známá informace bude vskutku odpovídat tomu, o čem je mluvčí přesvědčen, že adresát už ví.

Má-li počítačový systém korektně identifikovat známou informaci, měl by také testovat, že tato informace je konzistentní se základní bází znalostí, což mu umožní řešit případné víceznačnosti. Navíc, je-li jasně vyznačena i nová informace, systém může reagovat tím, že si ji doplní do svého modelu světa. Výraz *mu* tedy označoval mého studenta *Petra Nováka* a v tomto ohledu sotva může vzniknout nějaká nejednoznačnost.

Opozice určitosti–neurčitosti je často vodítkem pro rozlišení nového a známého. Např. ve větě

(v34) *Ten profesor zkoušel nějakého studenta.*

očekáváme (není-li k dispozici další kontext, který by naznačoval něco jiného), že *ten profesor* byl již zmíněn v konverzaci nebo je znám z kontextu, zatímco *nějaký student* se v konverzaci objevuje poprvé. Ne vždy je však situace tak jednoznačná a podobné jmenné skupiny často nesou i novou informaci. Jako příklad lze uvést

(v35) *Petr Novák nebyl včera ve škole. Tento úspěšný student orientující se na otázky českého slovosledu mě na dnešním semináři zklamal.*, kde vyznačená jmenná skupina jednak odkazuje k již zmíněné osobě a jednak o ní přináší novou informaci.

Dovedeme-li ve větě identifikovat známou informaci, můžeme jí využít k omezení množiny možných referentů u jmenných skupin. Ve větě o Petrovi (v33), kterému profesor vysvětloval pravidla českého slovosledu, dativní pronominální skupina musí odkazovat k někomu, kdo je student. Kdyby daná jmenná skupina byla víceznačná, mělo by smysl zjistit si implicitní presupozice potřebné k tomu, aby nová informace dávala smysl. Např. by nebylo konzistentní říci, že je něco lokomotiva, bylo-li již známo, že to je člověk nebo robot; podobně by bylo poněkud nekoherentní tvrdit o někom, že je předseda vlády, víme-li již, že jde o vysokoškolského studenta. Presupozice tohoto druhu lze ověřovat dopřednými inferenčními pravidly, která testují výskyt kontradikcí, např.:

dopravní-prostředek(X) if lokomotiva(X)
 počet-nohou(X,2) if člověk(X)
 počet-nohou(X,2) if robot(X)
 kontradikce if dopravní-prostředek(X) & počet-nohou(X,2)
 různé(X,Y) if uvnitř(X,Y)
 kontradikce if různé(X,X)

Tato pravidla nám bezprostředně pomohou odhalit kontradikci, když se pokusíme zpracovat následující zájmené referenty:

Robot předváděl nového Jaguára.

Byl *to* automobil.

(* “to” → robot)

Robot má dvě nohy.

Je *to* student.

(* “to” → robot)

Další způsob, jak testovat takové presupozice, představují zpětné inference. Kdykoli se chystáme doplnit do znalostní báze nějakou novou informaci, musíme testovat, zda je konzistentní s tím, co je již známo (uloženo v bázi). Postačující zpětná pravidla konzistence by mohla mít např. následující podobu:

konzistentní (lokomotiva(X)) if (dopravní-prostředek(X)) & ...

konzistentní (uvnitř(X,Y)) if (různé(X,Y)) & ...

Pravidla pro testování konzistence musejí přihlížet k pořadí, v němž bude pravděpodobně přicházet informace o objektech komunikace. Je možné, že zjistíme, jakého druhu objekt je, např. že to je dopravní prostředek, se dovíme dříve, než že jde o lokomotivu. Pak můžeme uplatnit předchozí pravidla, která potvrzují konzistenci – lokomotiva je typem dopravního prostředku.

Dovíme-li se však, že objekt je lokomotiva dříve, než je známo, o jaký typ objektu jde, pak první pravidlo ke stanovení konzistence nepostačuje. Naše pravidla konzistence (významové postuláty) ve skutečnosti nevyjadřují generalizace o světě, ale jsou to heuristická metalogická pravidla pro speciální případy, u nichž je nepravděpodobné, že by nová informace byla v kontradikci s tím, co je již známo. Smysl jejich použití je v tom, že mohou produkovat kandidáty na referenty, a méně již vést k zamítnutí referentů nevhodných či vysloveně chybných.

Efektivnější ovšem je snažit se přímo vydedukovat (najít) množinu propozic, které by měly být pravdivé, aby daná věta dávala smysl. Pak můžeme zamítnout nebo nepreferovat možné interpretace, které nepodporují pravdivost těchto propozic.

V praxi se často vyskytují situace, kdy formulace presupozic umožňuje adresátovi přímo provádět jednoduché inference, např. :

Marie má dvě děti, kluka a holku.

Dcera bude letos maturovat.

Můj kamarád koupil auto z druhé ruky.

Motor je v dobrém stavu, ale karosérie je shnilá.

Vhodná inferenční pravidla, která by měla být součástí našeho modelu porozumění jazyku, by mohla vypadat takto:

$dcera(X)$ if $dcera(X,Y)$

$dcera(X,Y)$ if $děvče(X)$ & $dítě(X,Y)$

$motor(motor(X))$ if $dopr.-prostředek(X)$

$karosérie(karosérie(X))$ if $dopr.-prostředek(X)$

Pravidla tohoto typu umožňují učinit závěr, že je-li $dopr.-prostředek$ 17 dopravní prostředek, pak existuje objekt $motor(dopr.-prostředek17)$, který je motorem dopravního prostředku. Užito dopředu vytvoří toto pravidlo automaticky objekt – motor, kdykoli se na scéně objeví dopravní prostředky. Při zpětné inferenci uvede na scénu motory dopravních prostředků tak, aby cíl inference byl splněn.

Prezentovaný pohled na distinkci známé (téma) – nové (réma) vychází, jak patrně, především z pozice porozumění přirozenému jazyku. S problémy podobného typu se ovšem musí vypořádat i jazykový generátor, u něhož je potřeba, aby explicitně poskytoval dostatečné množství tematických prvků (formálně signalizovaných osobními a ukazovacími zájmeny, koncovkami verba finita – povšimněme si tu zajímavé koincidence – zmíněné prvky hrají dvojí roli: signalizují téma a současně hrají svou roli deiktickou –, částicemi a některými dalšími prostředky), takže nebude docházet k chybnému přiřazování mezi příslušnými výrazy a jim odpovídajícími referenty.

0.8 Pragmatická rovina

Podrobná analýza vět přirozeného jazyka přesvědčivě ukazuje, že ani detailní sémantická analýza vět PJ, jak byla naznačena výše, nevyčerpává ještě plně problém porozumění větám PJ. Věty lze dále zkoumat z hlediska uživatele jazyka a z hlediska postojů, které uživatel (dále UJ) může zaujímat k sémantickému jádru věty, jímž pro nás, jak jsme už naznačili, je konstrukce + funkce konstrukcí konstruovaná. Zkoumání těchto otázek konstituuje pro nás oblast, kterou budeme dále nazývat **interní (vnitřní) pragmatika**.

I když přihlédneme k postojům UJ, i tak značná část vět PJ ještě nebude umožňovat jednoznačnou sémantickou interpretaci, pokud navíc nebudeme respektovat skutečnost, že vět se užívá v konkrétních **komunikačních situacích a kontextech**. Samotná sémantická analýza ukazuje, že mnohé věty jsou sémanticky neurčité, neboť neoznačují určitou konkrétní konstrukci, jak bychom očekávali, nýbrž nějakou **otevřenou konstrukci**. Zkoumání tohoto okruhu problémů konstituuje pro nás **externí (vnější) pragmatiku**.

0.8.1 Interní pragmatika

Ukázali jsme výše, že z hlediska sémantiky věta vyjadřuje konstrukci a denotuje propozici. Taková analýza ještě není úplná a snadno se lze přesvědčit o tom, že věta obsahuje ještě další informaci, která se týká UJ. Ve větě vždy najdeme specifické formální prostředky, které signalizují, že:

1. UJ pokládá propozici, kterou daná věta označuje, za **pravdivou** v nějakém (obvykle aktuálním) světě W a okamžiku S , pak jde o **tvrzení** formálně signalizované např. indikativem,
2. UJ chce zjistit, jaká je **pravdivostní hodnota** dané propozice – pak jde o empirickou **otázku**, a to buď o otázku zjišťovací, nebo o otázku doplňovací,
3. UJ chce, aby propozice odpovídající dané větě byla v aktuálním světě a okamžiku S pravdivá – potom jde o **rozkaz** formálně signalizovaný imperativem,
4. UJ si **přeje**, aby propozice odpovídající dané větě byla pravdivá v aktuálním světě a okamžiku S – pak jde o **přání**.

Můžeme tedy říci, že mimo to, co vyjadřuje a označuje, věta **demonstruje** uvedené postoje UJ. Soubor demonstrovaných postojů tvoří to, co bychom mohli nazvat **prostor postojů**.

Výše uvedené postoje představují **široké modality**, tj. postoje které mohou být demonstrovány ve větách deklarativních, interogativních, imperativních, deziderativních a dalších (např. typu nabídky, slibu, odmítnutí).

Dalším druhem postojů jsou **jistotní modality**, tj. postoje demonstrující subjektivní míru pravděpodobnosti toho, že daná propozice v aktuálním světě a okamžiku S platí.

Formálními prostředky tu jsou modální slovesa (*mušet, moci, mít*) a modální adverbia a částice typu *asi, snad, možná, jistě, určitě*. Lze uvažovat ještě o dalších druzích postojů, jak jsou naznačeny např. v práci Materna, Pala, Svoboda, 1979.

0.8.2 Externí pragmatika

Výsledkem sémantické analýzy vět jsou často tzv. **otevřené konstrukce**, tj. konstrukce, v nichž se vyskytují volné proměnné. V takových případech sémantická analýza nedostačuje k určení, o kterou konkrétní propozici jde, a proto je nutno přejít k analýze pragmatické. Otevřené konstrukce odpovídají vždy nějaké třídě propozic – jsou tudíž víceznačné. Volné proměnné se v konstrukcích objevují zpravidla tam, kde se v odpovídajících analyzovaných větách vyskytly výrazy v literatuře charakterizované jako **deiktické (indexové)**. Patří k nim např. osobní zájmena *já, ty, on, my, ...*, ukazovací zájmena *ten, ta, to, tehle, tamten, ...*, místní adverbia *zde, tady, tam, ...*

Deiktické výrazy odkazují ke komunikační situaci, v níž je příslušná věta proslouvena. Komunikační situace umožňuje určit, jaké konkrétní atomy (konstanty) mají být dosazeny za volné proměnné získané v průběhu sémantické analýzy při budování SR analyzované věty. Teprve tak získáme **uzavřené konstrukce**, jež konstruují konkrétní propozice.

Komunikační situaci můžeme charakterizovat jako vektor $(t, l, m, h, o_1, \dots, o_n)$, kde

t – je časový okamžik

l – je nějaké místo (prostor)

m – je mluvčí

h – je posluchač

o_1, \dots, o_n – jsou objekty univerza, o nichž se právě (v dané větě) mluví.

Pro jednotlivé složky věty

(v36) *Ona je studentka.*

nechť máme v sémantickém slovníku následující typy:

být studentkou **S/** $(o_l)_{\tau\omega}$ – vlastnost individuí

ona x/l – proměnná individuí

Větě (v36) pak odpovídá otevřená konstrukce

(K4) $\lambda w \lambda t (S_{wt}(x))$.

Abychom zjistili, která konkrétní propozice je konstrukcí (K4) konstruována, musíme vzít v potaz konkrétní komunikační situaci KS_3 , jež určuje, kdo je individuum, o němž se mluví ve (v36).

Lze to učinit pomocí **pragmatické funkce** F_{ona} , jejímž oborem je množina komunikačních situací. Funkce F_{ona} určuje, jaká valuace má být vybrána pro větu (v36).

Konstrukci (K4) můžeme s použitím funkce F_{ona} zapsat následujícím způsobem:

(K5) $\lambda w \lambda t (\mathbf{S}_{wt}(x[F_{ona}]))$.

Jestliže se v situaci KS mluví o individuu \mathbf{AN} , je $F_{ona}(S) = \mathbf{AN}$ a konstrukce (K5) pak vypadá takto:

(K6) $\lambda w \lambda t (\mathbf{S}_{wt}(\mathbf{AN}))$.

Ta již je uzavřená a konstruuje konkrétní propozici, jíž odpovídá např. věta (v36a) *Alena Nováková je studentka*.

Tím jsme naznačili jeden možný průběh pragmatické analýzy vět, jako je (v36), v rámci externí pragmatiky, neodpověděli jsme tím však ještě na otázku, jak obecně budovat pragmatické funkce, tj. jak obecně budovat algoritmus přechodu od sémantiky k externí pragmatice.

Pokusme se aspoň stručně nastínit, jak by se v tomto směru dalo postupovat s ohledem na systémy pro porozumění přirozenému jazyku. V každém případě se lze opírat o deiktické výrazy a už při syntaktické a sémantické analýze se pokusit o vymezení komunikační situace jako celku. K tomu je potřeba určit hodnoty jednotlivých proměnných konstituujících komunikační situaci jako celek, tj.:

1. nalézt nebo stanovit hodnotu proměnné t , což může spočívat ve zjištění nebo zadání daného data včetně konkrétního časového okamžiku – zde jsou východiskem gramatické časy a další časové výrazy, ostatně všechny počítačové systémy (operační systémy zejména) jsou dnes vybaveny hodinami a kalendářem, takže potřebné informace o čase dané komunikace mohou být snadno k dispozici,
2. určit hodnotu proměnné l , tedy explicitně identifikovat místo, na němž daná komunikace probíhá. Na rozdíl od časových údajů není tato informace vyjadřována gramatickými prostředky, ale jen lexikálně jistými typy adverbíí, případně dalšími výrazy. V současných počítačových systémech není informace o místě pokládána za relevantní, nicméně pro komunikaci v přirozeném jazyce bude nevyhnutelné s ní počítat,
3. identifikovat hodnoty proměnných m a h , tj. zjistit, kdo je v dané komunikační situaci *mluvčím* a kdo *posluchačem* a jaký mají vztah k objektům o_1, \dots, o_n , což je spolehlivě signalizováno prostředky vyjadřujícími gramatické osoby (osobní zájmena a koncovky verba finita),
4. určit, o kterých objektech univerza jde v dané promluvě řeč, znamená nalézt jejich referenci, tj. provést sémantickou analýzu dané promluvy. Tento krok je úzce spojen s přechozími body, ale na tomto místě je obtížné stanovit posloupnost jednotlivých akcí, které povedou nejen k získání sémantické reprezentace dané výpovědi, ale také zajistí provázání s komunikační situací, i když je zřejmé, že nejnadějnější řešení by mělo směřovat k paralelnímu zpracovávání předchozích tří bodů.

0.9 Dialogové systémy, inference

0.9.1 Analýza promluvy, promluvvé objekty

0.9.2 Anafora, anaforické vztahy

0.9.3 Odkazovací výrazy, rozpoznávání antecedentů

0.9.4 Historie promluvy a promluvvý zásobník

0.9.5 Segmenty v promluvě

0.10 Závěr

Pokusme se shrnout výše uvedené výsledky. V oblasti české morfologie se nám podařilo vytvořit algoritmický popis české deklinace a konjugace pokrývající odhadem 80 % české slovní zásoby – náš současný slovník českých kmenů kmenů čítá něco přes 170 000 položek. V algoritmickém popisu se dále propracovává systém vzorů, zejména u sloves dochází k propojení vzorů s prefixy včetně začlenění popisu vidů, což vede k výraznému zpřehlednění této části popisu zahrnující asi 70 000 českých sloves a také k jeho další optimalizaci (zkrácení o více než 50 %). K dispozici již je první verze lemmatizátoru, který byl začleněn do první varianty počítačového synonymického slovníku češtiny (v rozsahu kolem 20 000 hesel) a po dokončení potřebných úprav bude existovat i jako samostatný modul použitelný např. v rešeršních systémech a dalších vhodných aplikacích. Práce na algoritmickém popisu bude dále pokračovat zejména v oblasti slovotvorby, v níž bychom rádi dospěli k vytvoření slovotvorného automatu, tj. programu, který by modeloval hlavní slovotvorné procesy v češtině a měl by schopnost interaktivně se učit.

Jak jsme ukázali v další části práce, využili jsme příznivých vlastností PROLOGU a v programu KLARA naznačili integraci algoritmického popisu morfologie a syntaxe. V programu KLARA II je pak tento postup ilustrován na českých slovesech označujících komunikaci a je ho využito i pro vytvoření jednoduchého, avšak dostatečně zajímavého programu překládajícího věty se slovesy komunikace z češtiny do angličtiny. Naším nejbližším cílem v tomto ohledu je pokusit se o integraci české morfologie a syntaxe na kvalitativně vyšší úrovni dané velkým rozsahem slovníku, s nímž je již schopen pracovat morfologický analyzátor, a vytvořit syntaktický analyzátor (generátor) schopný pracovat se souvislými českými texty (v aplikaci použitelný např. jako gramatický korektor).

Pokud jde o rovinu sémantickou, využili jsme dřívějších výsledků a pokusili jsme se naznačit jednu z možných cest, která může vést k integraci syntaxe a sémantiky a posléze i pragmatiky. Zde prezentovaný přístup se v daném okamžiku pohybuje více v oblasti teoretického hledání než přímých počítačově orientovaných aplikací, i když

v dílčích úsecích jsou již docela dobře možné. Ukazuje se, že při práci na integraci morfologie a syntaxe bude vhodné a potřebné orientovat se současně i na začlenění sémantiky do takto naznačeného analyzátoru. Stejně tak je zřejmé, že v oblasti sémantiky se neobejdeme bez nemalé práce empirické, která se týká jednak otázek lexikálních včetně získávání dat ze strojově čitelných slovníků a jednak sémantické analýzy víceslovných výrazů a vět s využitím TILU.

V tomto bodě citelně pociťujeme nedostatek vhodného a uživatelsky „přítulnějšího“ programového vybavení pro práci s gramatikami a reprezentacemi znalostí, které by umožnilo zajímavé a k dalšímu poznání vedoucí experimenty v naznačené oblasti. Nevyhnutelná je jak těsná spolupráce s kvalitními odborníky v oblasti počítačové vědy a AI, tak i kvalitní technické vybavení, což je v současnosti především záležitost dostatečných finančních prostředků.

Literatura

- Akademická mluvnice češtiny, ed. Petr, J., kol. autorů, Mluvnice češtiny 1,2,3, Praha 1986.
- Benešovský, M., Šmídek, M., Testování programů, sb. semináře SOFSEM 1984, VUSEIAR Bratislava, 1984.
- Bierwisch, M., Strukturelle Semantik, in: Deutsch als Fremdesprache 6, Heft 2, s.67, 1969.
- Clocksin, W., Mellish, Ch., Programming in PROLOG, Springer-Verlag, Berlin, 1981.
- Colmerauer, A., Metamorphosis grammars, in: *Natural Language Communication with Computers*, ed. L. Bolc, Springer Verlag, s.133-89, 1978.
- Čermák, F., Králík, J., Pala, K., Počítačová lexikografie a čeština (*Počítačový fond češtiny*), Slovo a slovesnost, 53, 41-48, 1992.
- Čermák, F., Holub, J., Syntagmatika a paradigmatica českého slova I (Valence a kolokabilita), skriptum LŠSS, UK Karolinum, Praha 1991.
- Čihánek, P., Sémantický analyzátor pro češtinu, rigorózní práce, Brno 1978.
- Dahl, V., Abramson, H., On gapping grammars, in: Proceedings of the Second Int. Conference on Logic Programming, Ord & Form, Uppsala, Sweden, s.77-88, July 1984.
- Daneš, F., Hlavsa, Z., Větné vzorce v češtině, Academia, Praha, 1981.
- Dokulil, M., Daneš, F., *K tzv. významové a mluvnické stavbě věty*, in: O vědeckém poznání soudobých jazyků, Praha, s.231-246, 1958.
- Fillmore, Ch., J., The case for case, in: *Universals in Linguistic Theory*, E. Bach and R. Harms, eds., Holt, Rinehart & Winston, New York, s.1-88, 1968.
- Firbas, J., *On the Concept of Communicative Dynamism in the Theory of FSP*, SBP-FFBU, A 19, Brno, s.135-144, 1971.
- Frege, G., Über Sinn und Bedeutung, in: Zeitschrift für Philosophie un philosophische Kritik (Halle) 1892, NF 100, s.25-50.
- Gazdar, G., Mellish, Ch., Natural Language Processing in: PROLOG, Addison Wesley,, Wokingham, 1989.
- Grepl, M., Karlík, P., *Skladba spisovné češtiny*, SPN, Praha, 1987.
- Grosz, B., J., The representation and use of focus in dialogue understanding, PhD. dissertation, University of California at Berkeley, 1977.

- Hajič, J., Drozd, J., Spelling-Checking for Highly Inflected Languages, sb. konference COLING'90, Helsinki, 1990.
- Hajičová, E., Sgall, P., Towards an automatic identification of topic and focus, *ACL Proceedings, Second European Conference*, s.263-7, 1985.
- Havránek, B., Jedlička, A., *Česká mluvnice*, Academia, Praha, 1960.
- Church, A., Introduction to mathematical logic, Princeton 1956.
- Katz, J., J., Fodor, J., A., The structure of a semantic theory, *Language* 39, 1963, 170-210.
- Komárek, M., Ke dvěma koncepcím stavby jednoduchých slovesných tvarů v češtině. *Acta Universitatis Palackianae Olomucensis. Studia Bohemica IV*. Praha 1987.
- Konečná, D., Algoritmické popisy českých slovesných tvarů, disertační práce, FF UK Praha, 1964.
- Koskenniemi, A general computational model for word form recognition and production, COLING-84, s.178-81, 1984.
- Kulagina, O., S., Mel'čuk, I., A., Mašinnyj perevod s francuzskogo jazyka na russkij, *Voprosy jazykoznanija* 5, Moskva, 1956.
- Machová, S., Havel, I., M., Pala, K., Komunikace s počítačem v přirozeném jazyce, *Materiály semináře SOFSEM 1978*, VUSEIAR Bratislava, 1978.
- Machová, S., Říha, A., Computer testing of generative grammar, *PBML* 29, Praha, s.43-58, 1978.
- Materna, P., An Intensional approach to questions, *Kybernetika* 15, s.161-192, 1979.
- Materna, P., Pala, K., *Theoretical framework for syntax and semantics*, Sborník celostátní konference o kybernetice, Praha, 1976.
- Materna, P., Pala, K., Svoboda, A., Externí a interní pragmatika, *Otázky slovanské syntaxe IV/1*, 53-60, Brno, 1976.
- Materna, P., Pala, K., Svoboda, A., The ordered-triple theory continued, *Brno Studies in English* 13, 119-165, 1979.
- Materna, P., Sgall, P., Hajičová, E., „Linguistic constructions“ in transparent intensional logic, in: *Categorial Grammar*, ed. by W. Buszkowski, W. Marciszewski and J. van Benthem, John Benjamins Publishing Co., Amsterdam/Philadelphia, s.283-300, 1988.
- Mel'čuk, I., A., *Avtomatičeskij sintaksičeskij analiz*, Novosibirsk, 1964.

- Minsky, M., A framework for representing knowledge, in: *Mind Design*, ed. J. Haugeland, MIT Press, Cambridge, 95-128, 1981.
- Montague, R., *Formal Philosophy*, ed. by R. H. Thomason, Yale University Press, New Haven and London, 1974.
- Osolsobě, K., *Algoritmický popis české formální morfologie substantiv a adjektiv*, rukopis pro SBPFFBU, Brno 1988.
- Osolsobě, K., *Model vybraných slovotvorných typů (v jazyce PROLOG)*, rukopis, Brno 1990.
- Osolsobě, K., *Popis systému českých substantivních a slovesných vzorů*, rukopis disertační práce, Brno, 1991.
- Osolsobě, K., Pala, K., *Czech Stem Dictionary for IBM PC XT/AT*, Conference on Computer Lexicography, Balatonfüred, September 1990.
- Osolsobě, K., Pala, K., *Základy počítačové lingvistiky*, vš. skriptum, FF MU, Brno 1992.
- Pala, K., *O procedurální gramatice (pro češtinu)*, SBPFFBU, A 30, 103-122, Brno 1982.
- Pala, K., *O sémantických reprezentacích*, SBPFFBU, A 32, 24-35, Brno 1984.
- Pala, K., Osolsobě, Franc, S., *Česká morfologie a syntax v PROLOGU*, SOFSEM 1987, VUSEIAR. Bratislava 1987.
- Páleš, E., *SAPFO – systém pre komunikáciu v prirodzenom jazyku*, dipl. práce, MFF UK, Bratislava, 1988.
- Palová-Vaničková, I., *Syntaktický analyzátor pro češtinu*, rigorózní práce, Brno 1977.
- Panevová, J., *Random generation of Czech Sentences*, Proceedings of COLING 82, ed. by J. Horecký, Academia, Praha 1982.
- Panevová, J., *Verbal frames revisited*, PBML 28, s.55-72, 1978.
- Pereira, Fernando, C., N., Warren, David, H., D., 1980, *Definite clause grammars for language analysis – a survey of the formalism and a comparison with ATN*, *Artificial Intelligence*, 13, 231-78.
- Piřha, P., *On the case frames of nouns*, PSML 7, Academia, Praha, s.215-224, 1981.
- Podlezlová-Koželouhová, B., *Sémanticky orientovaný generativní popis českých sloves nepřechodných*, diplomová práce, FF MU Brno, 1974.

- Quillian, M., R., Semantic memory, in: *Semantic Information Processing*, ed. by M. Minsky, MIT Press, Cambridge, Mass., s.227-270, 1968.
- Sgall, P., Soustava pádových koncovek v češtině, *AUC – Slavica Pragensia* 2, s.65-84, 1960.
- Sgall, P., Generativní popis jazyka a česká deklinace, Academia, Praha 1967.
- Sgall, P., a kol., Úvod do syntaxe a sémantiky, Academia, Praha, 1985, s.9.
- Sgall, P., et al, The Meaning of the sentence in its semantic and pragmatic aspects, Academia, Prague, 1986,
- Sgall, P., Hajičová, E., Buráňová, E., Aktuální členění věty v češtině, Academia, Praha, 1980.
- Schank, R., Conceptual dependency: a theory of natural language understanding, *Cognitive Psychology*, 3, 552-631, 1972.
- Slovník spisovného jazyka českého, Academia, Praha, 1960, 1989.
- Ševeček, P., Morfologické programy pro češtinu: analyzátor a lemmatizátor, rkp., 1992.
- Šmilauer, Vl., *Novočeská skladba*, SPN, Praha, 1969.
- Tichý, P., Introduction to intensional logic, rukopis, University of Otago, 1976.
- Tichý, P., The Semantic of episodic verbs, *Theoretical Linguistic* 7, s.263-296, 1980.
- Tichý, P., The foundations of Frege's Logic, de Gruyter, Berlin – New York, 1988.
- Wampler, B., E., and the RSI Software Engineering Staff, GRAMMATIK IV, v. 1, Software International, 1989.
- Winograd, T., *Understanding Natural Language*, Academic Press, New York, 1972.
- Woods, W., 1973, Progress in natural language understanding: an application to lunar geology, *AFIPS Conference Proceedings*, 42, 441-50.
- Osolsobě, K., Algoritmický popis české formální morfologie, disertační práce, Brno 1996.
- Panevová, J., On Verbal Frames in Functional Generative Description, Part I, II, *The Prague Bulletin of Mathematical Linguistics* 22, pp.3-39.
- Pala, K., Všianský J., Slovník českých synonym, NLN Praha, 1995,

Petr, J., a kol., Mluvnice češtiny I, II, Academia Praha, 1986,

Slovník spisovného jazyka českého, Academia Praha, 1.vyd. 1960, 2.vyd. 1989

Somers, H., L., Valency and Case in Computational Linguistics, eds. S. Michaelson and Y. Wilks, Edinburgh Information Technology Series, Edinburgh University Press, 1987, pp.4-29

Svozilová N. a kol. Valenční slovník vybraných českých sloves, ÚJČ ČAV, Praha, 1997
???

Ševeček, P., Morfologický analyzátor a lemmatizátor pro češtinu – implementace v jazyce C, program (rukopis), Brno, 1995